Lethal-trifecta enforcement

The one detector that is about a *combination* rather than a *thing*. **Default: warn.**

The idea

An AI agent becomes dangerous when three capabilities meet in the same context:

1. **Access to private data** — your repository, your database, your issue tracker.
2. **Exposure to untrusted content** — a web page, an email, an issue written by someone outside your company.
3. **A way to send data out** — an HTTP tool, an email tool, a webhook.

Any one is fine. Any two are usually fine. All three together means a malicious instruction hidden in the untrusted content can direct the agent to read your private data and send it somewhere. The agent is not compromised in any traditional sense; it is doing exactly what it was asked, by the wrong person.

What Sentilai does

Per tool call, per device, the Gateway tracks which of the three legs the device's current context has. When a call would complete the set, the detector's action applies.

MCP Inventory shows a banner when a device has all three legs present, naming the `server/tool` that contributed each one. That banner is the useful part even in warn mode: it tells you which combination of servers created the exposure, so you can decide which one does not belong.

Where it applies

On the MCP path — the local shim and routed remote endpoints — because that is where tool calls are visible with names. It is not a content scan of the prompt.

Setting it

Policy ? Risk detectors ? Lethal-trifecta enforcement. Warn first; look at the banners; then block once you know which combinations are real.

Revision #6
Created 2026-08-02 10:20:02 UTC by Sentilai Docs
Updated 2026-08-04 08:39:45 UTC by Sentilai Docs