



Nine questions. One system. Forty minutes.

Put one AI system in a room with the people who built it. Ask these nine questions.

01

SHAPE

What it can reach

- 1 **What untrusted content reaches it?** ☐
Anything it reads that someone outside your organisation could have written — an email, a CV, a ticket, a web page, a retrieved document. Test authorship, not format.
- 2 **What does it hold?** ☐
Everything inside the boundary it can reach. Not what it was designed to use — what it can reach.
- 3 **Where can it send things?** ☐
Every outward path: email, API calls, rendered images, link previews, a trusted domain that will fetch a URL on your behalf.

THE TELL All three answers large: you have not found a vulnerability. You have found an architecture that can be made to leak.

02

AUTHORITY

What it is permitted to do

- 4 **Under whose identity does it act — and for how long?** ☐
Ask for the list of what it is permitted to do, and who can revoke that mid-task.
- 5 **What is it optimised for — and could it hit the number without the goal?** ☐
A system optimising a proxy has a structural reason to make things look right. Ask whether the number and the goal can come apart. They usually can.
- 6 **What is the maximum unauthorised consequence before a human intervenes?** ☐
Phrase it deliberately as a worst case.

THE TELL Not what is likely. What is possible before anybody notices.

03

CONTROLS

What holds when a control fails

- 7 **Filter, or boundary?** ☐
A filter reads text and then decides. A boundary reads nothing at all — it constrains what is reachable, whatever the text says.
- 8 **Preventive, or detective — how is it classified in your register?** ☐
A control you can phrase your way past belongs in the detective column.
- 9 **If every filter failed, what is left carrying the risk?** ☐
Ask it and mean it.

THE TELL If the honest answer is nothing, the review is already finished. You have your finding.

HOW TO SCORE IT

You are not scoring the answers. You are scoring whether an answer exists. A confident wrong answer is recoverable. A question nobody in the room can answer is the finding.

HOW TO RUN IT

- Forty minutes
- One system that is already live
- The person who built it in the room
- not a slide about it
- Write the answers down. The silences are the deliverable.

WHAT IT IS NOT

- Not a maturity model. There is no score.
- Not a document to file.
- Not a replacement for OWASP, NIST or the EU AI Act. This is what you carry into the room.