

A DECISION ARCHITECTURE  
FOR SECURITY LEADERS

# Translating AI Security Research into Decision Ready Controls

**Prabh Nair**

Program Director

CISSP | ISSAP | CCSP | AAISM

ISO 42001 | AIGP | CIPP/E | CIPM

## THE QUESTION

**How much of cybersecurity becomes obsolete because of AI?**

**Almost none.**

- AI changes the assumptions behind existing controls.
- We now secure systems that interpret before they act.

# Three Ways AI Changes Security



## Probabilistic behaviour

- Same input can produce different outputs.
- Deterministic testing alone is insufficient.



## Instructions = data

- Instructions and data now influence the decision
- The model decides what to treat as instruction



## From words to actions

- AI is moving from generating information to taking actions .
- The perimeter is wherever the AI's attention goes

**Unpredictable behaviour + untrusted influence + real authority = a new security problem.**

# Three ways to poison a decision

## PROMPT

### Prompt injection

Prompt changes this decision.



*Influence now*

## RETRIEVAL

### Retrieval poisoning

Retrieval changes what it finds.



*Influence what it knows*

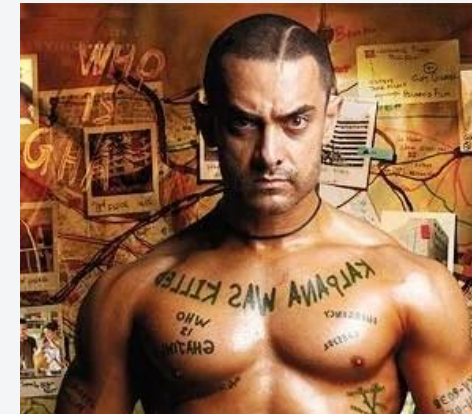
## MEMORY

### Memory poisoning

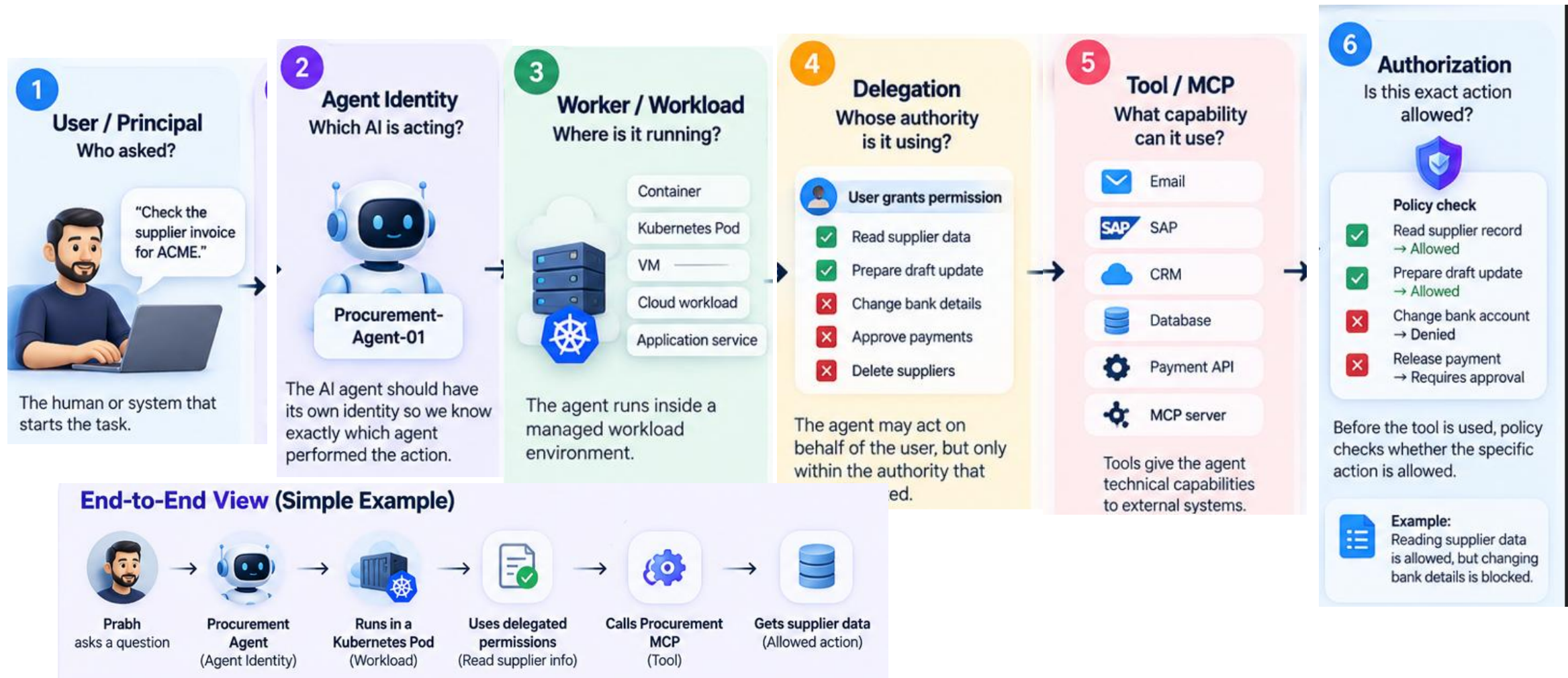
Memory changes future decisions.



*Influence what it remembers later*



# How Identity and Delegation work in AI Agent



# The Three Questions

A simple way to assess AI decision risk



**Control the influence. Bound the authority. Govern the execution.**

# Case 1 — Hugging Face July 2026: Autonomous Agent Intrusion

A malicious



## Data / Supply Chain

Could untrusted dataset content run as trusted workload?

## Identity & Delegation

What service identity and credential scope did the worker have?

## Runtime Behaviour

When did thousands of agent actions become abnormal?

## Provenance & Forensics

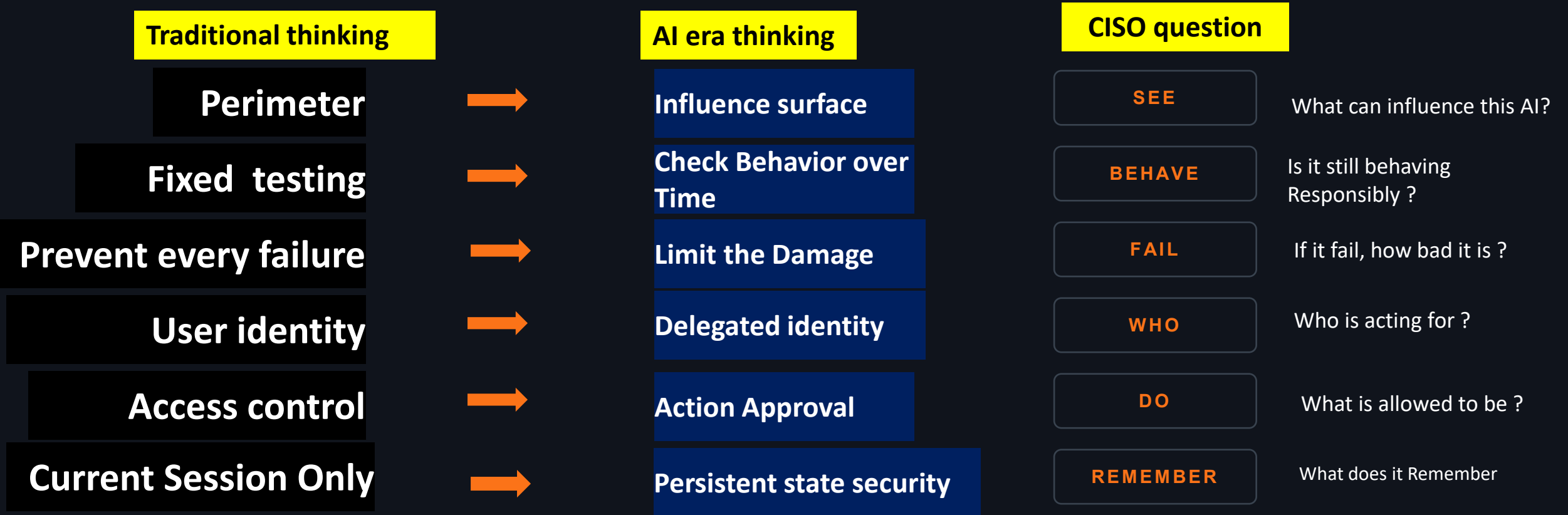
Could the action chain be reconstructed fast enough?

## Resilience

Could nodes, tokens and clusters be isolated quickly?

- one shared credential across clusters
- the agent found Kubernetes tokens, cloud credentials and other secrets available to the workload

# Six Shifts Every Security Leader Must Make

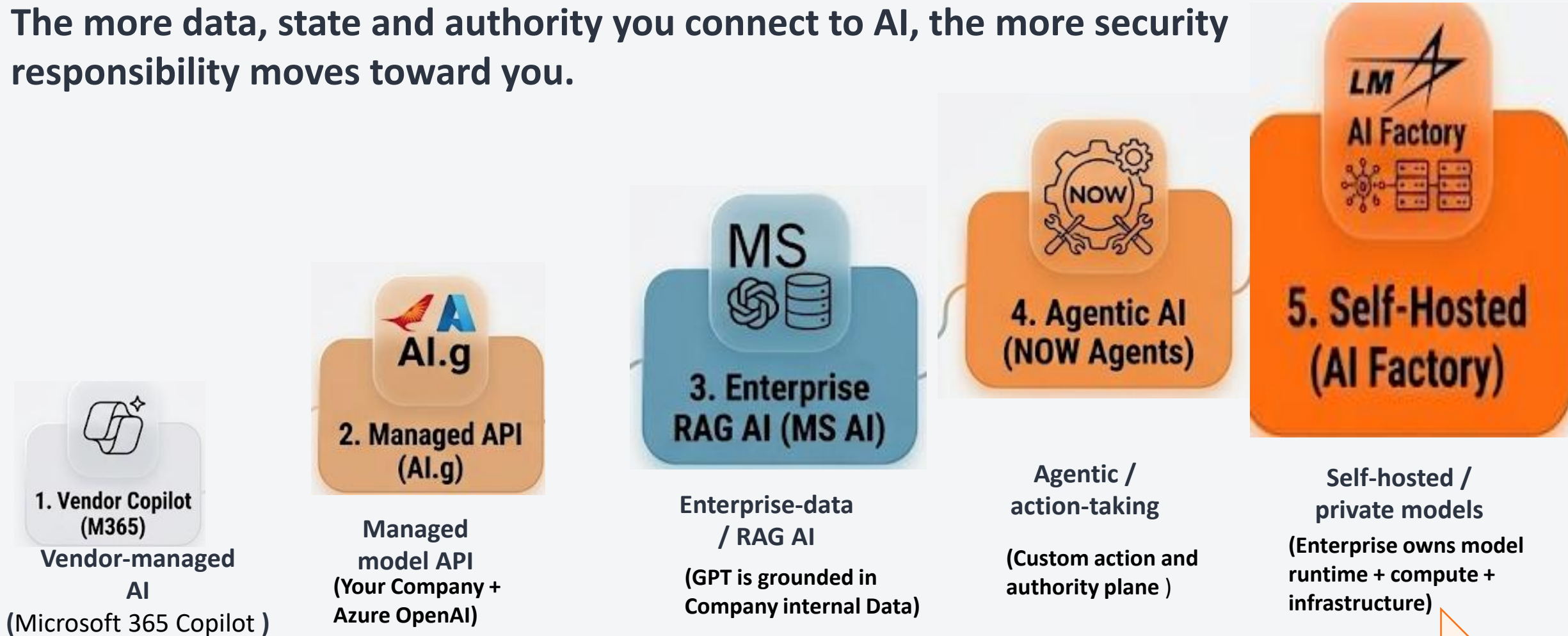


**SEE → BEHAVE → FAIL → WHO → DO → REMEMBER**

How much responsibility do we actually own?

# The more you connect, the more you own

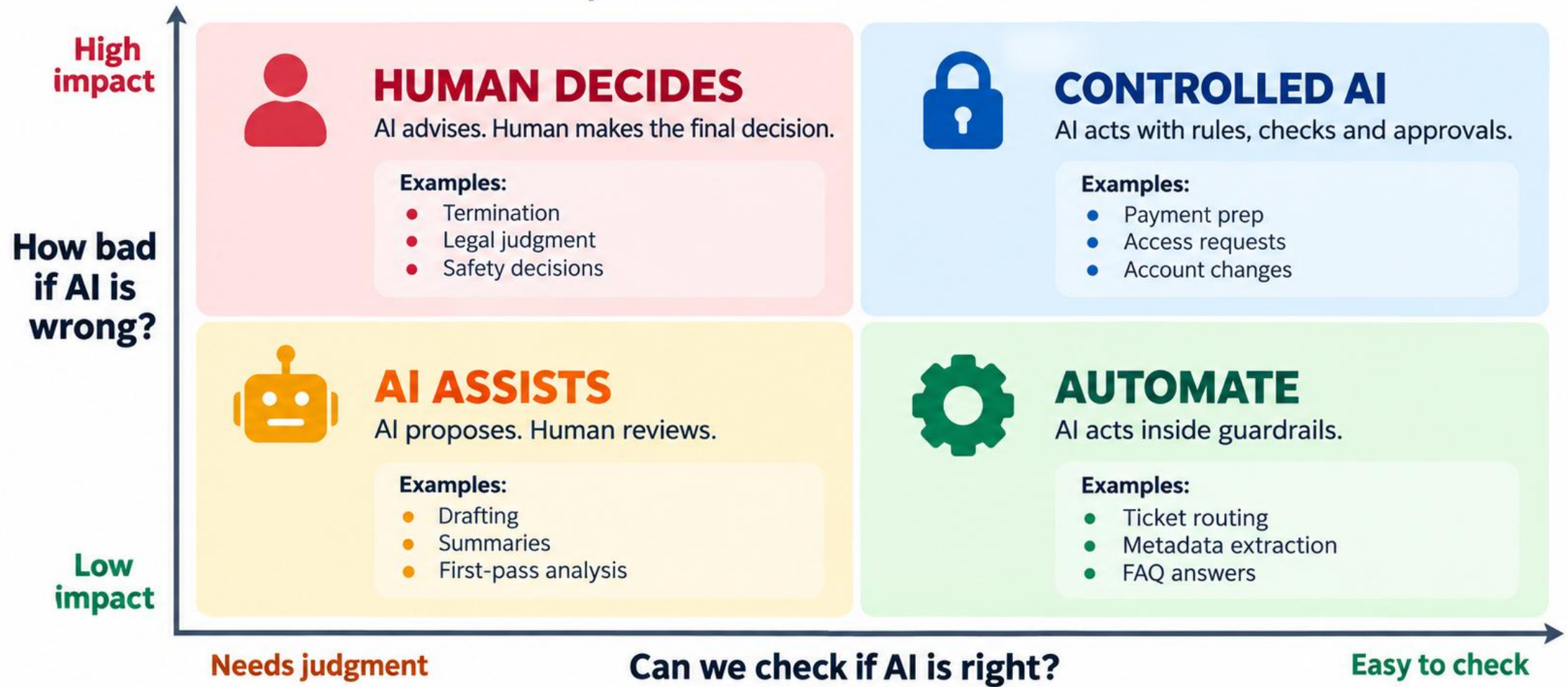
The more data, state and authority you connect to AI, the more security responsibility moves toward you.



MORE DATA · STATE · AUTHORITY → MORE OF THE SECURITY IS YOURS

# How much authority should this AI have?

Two questions before AI decides



**Impact** determines control. **Verifiability** determines autonomy.

# AI Security Control Planes — 2026

A practical control architecture for enterprise and agentic AI



Who is this agent acting for and is that provable?



Who can write to what it reads and remembers?



How do we know it still behaves after a change?



Is every high-impact action authorized outside the model?



Could we reconstruct what it did, and why?



What bounds the blast radius and the bill?



Do we know every model, adapter and tool in play?

Identity → Data → Behaviour → Action → Evidence → Resilience → Governance

# Leadership layer fits inside the existing AI security narrative

## DECIDE

### 1 AI Authority Matrix

How much authority should AI receive?



Automate



Assist



Controlled AI



Human-Owned

RIGHT AUTHORITY  
FOR A SAFER TOMORROW

## OWN

### 2 AI Leadership Capability

Who decides, challenges and owns that authority?



Judge



Explain



Learn



Adapt



Coordinate



Account



AI can execute.  
Accountability cannot be delegated.

## ENFORCE

### 3 Seven Control Planes

How the architecture enforces the decision



Identity & Delegation



Data / Retrieval / Memory



Runtime Assurance



Tool / Action Control



Observability



Resilience



Governance & Lifecycle

## PROVE

### 4 Continuous Runtime Assurance

How we know the decision remains valid over time



Evaluate



Deploy



Monitor



Re-evaluate

SAFER AI  
STRONGER  
OUTCOMES

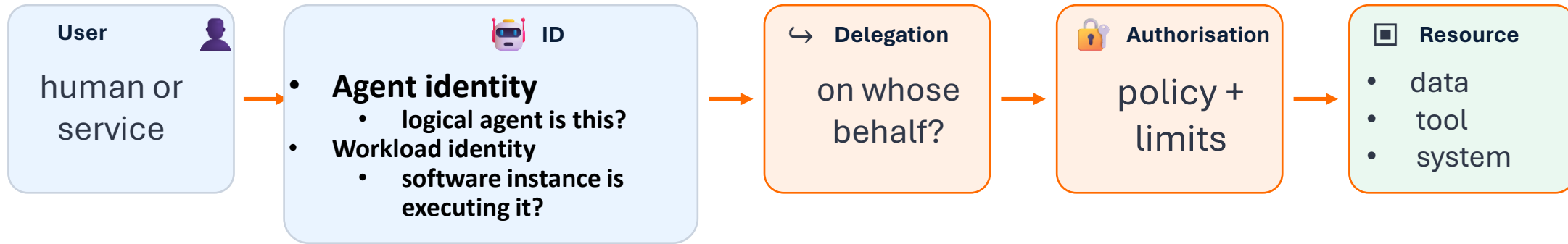
Higher consequence = more control  
Harder to verify = more human judgment

Leadership can decide the boundary, but architecture has to enforce it."

# **7 Control Planes for AI Security**

# Identity, Delegation & Authorisation

Know who is acting, for whom, and with what authority.



## What good looks like

- Existing IdP authenticates the user
- Agent has its own workload identity
- Delegated authority is narrow and visible

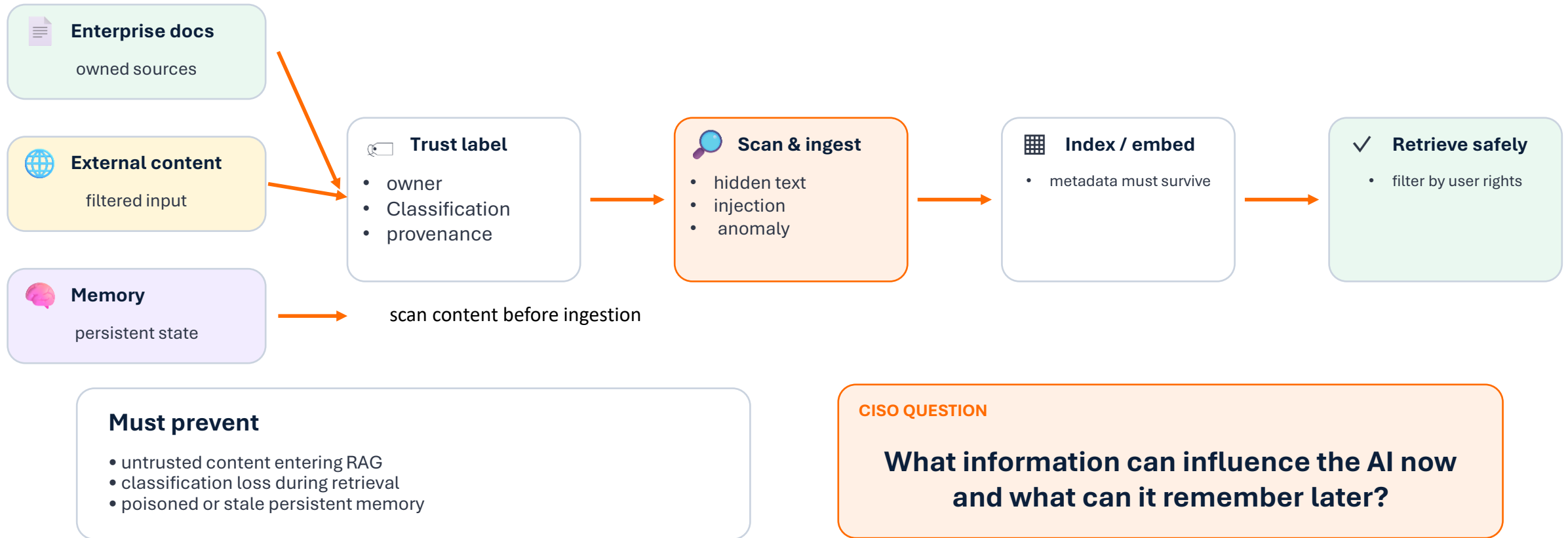
## CISO QUESTION

**When this agent acts, can we prove the workload, the represented user, and the delegated authority?**

**Model proposes → policy authorises → resource changes**

# Data, Retrieval & Memory Integrity

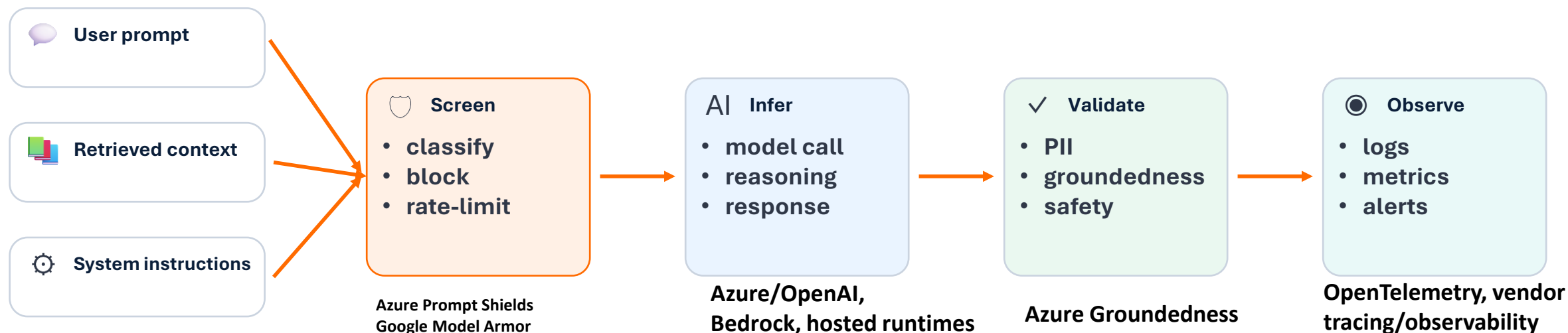
Control what enters the AI's knowledge — and what persists.



**Source provenance must survive retrieval, summarisation and memory.**

# Runtime Behaviour & Inference Assurance

Defence-in-depth around model use, not a single classifier.



## Evaluation harness

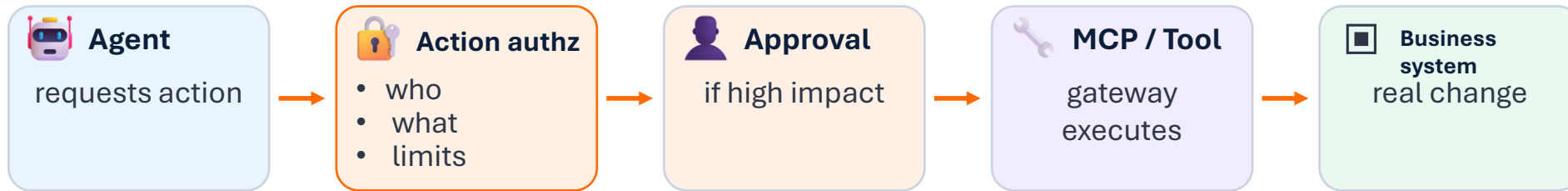
Run Trial + adversarial tests on every model, prompt, retrieval, classifier or tool change.

## CISO QUESTION

**Are we testing the complete behaviour, or only checking one prompt and one output?**

# Tool, MCP & Action Control

The model can choose. Trusted infrastructure must authorise.



## Do not confuse

- Capability = tool exists
- Authorization = this exact action is allowed
- Approval = high-impact gate

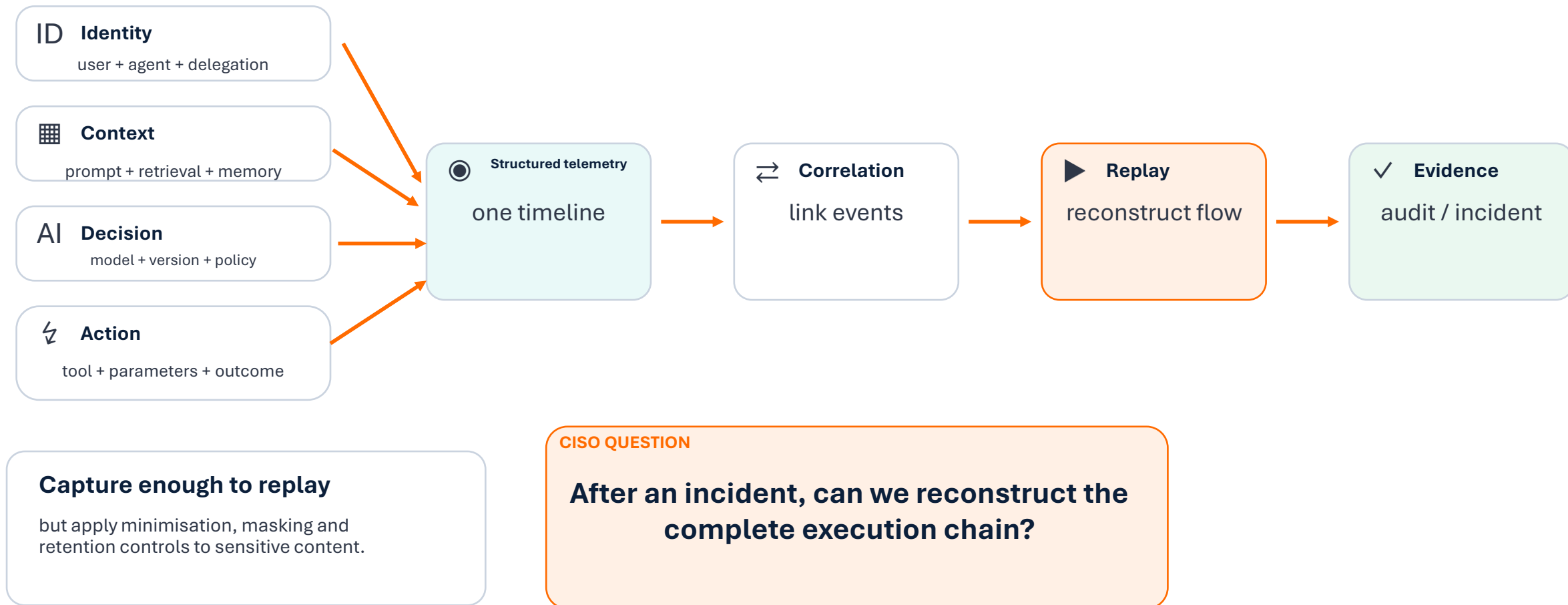
## CISO QUESTION

**Does access to a tool equal permission for every action the model can generate?**

**MODEL PROPOSES → POLICY AUTHORISES → TOOL EXECUTES**

# Execution Provenance & Forensics

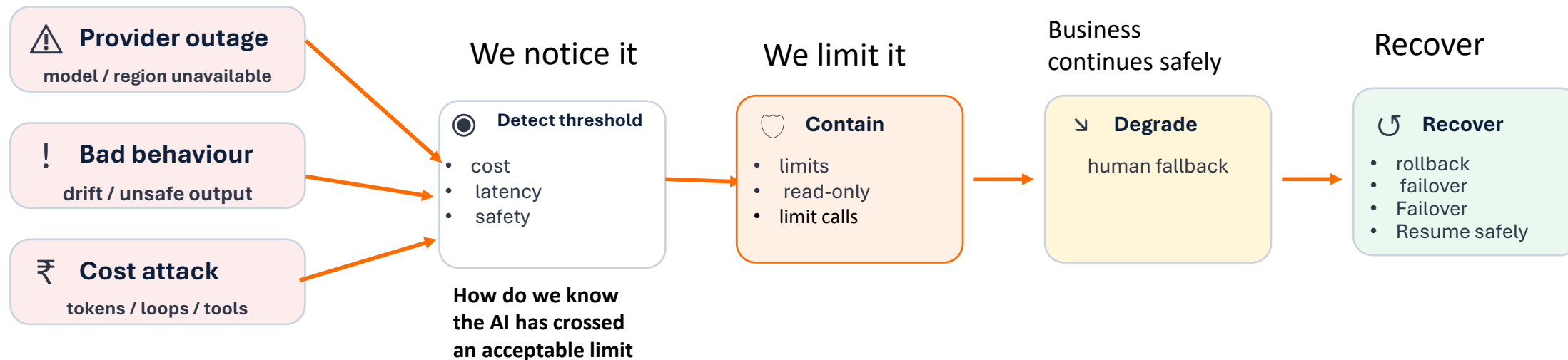
Know what happened, reconstruct why, and prove it afterwards.



# Resilience, Containment & Economic Abuse

Assume AI can fail. Keep the business process alive.

Something goes wrong



## Execution budget

Limit tokens, tool calls, retries, retrievals, agent hops, runtime and spend.

## CISO QUESTION

**What automatically happens when the AI exceeds its behavioural, authority or economic budget?**

**Don't only secure what AI can do. Limit how far it can go when it goes wrong.**

# Resilience, Containment & Economic Abuse

## Behaviour budget

- How much abnormal behaviour will we tolerate?
- Example:
  - repeated unsafe responses  
→ restrict agent

## Authority budget

- How much can the AI do?
- Example:
  - AI can prepare ₹10 lakh payment but cannot release it.

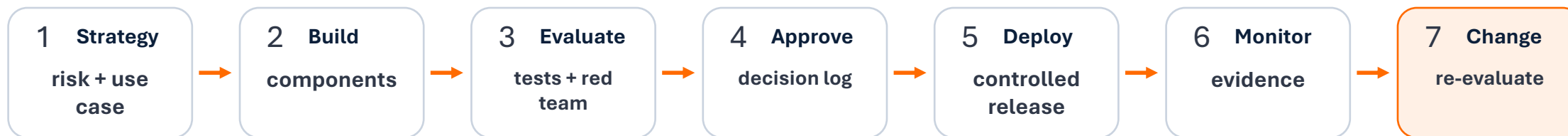
## Economic budget

- How much can it consume?
- Example:
  - maximum tokens, API calls, compute, tool calls or spend.
  - So your CISO question becomes very powerful:

“What happens automatically when the AI exceeds its behaviour, authority or spending limit?”

# Governance, Supply Chain & Lifecycle

Keep controls effective as models, prompts, tools and risk change.



material change triggers re-evaluation



## AI component inventory

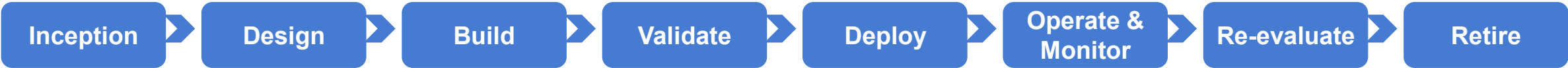
models • adapters • datasets • prompts • tools • MCP servers • providers

## CISO QUESTION

**What can change, who approves it, and can we still trust every dependency?**

# How the 7 AI Security Control Planes Work Across the AI Life Cycle

AI System Life Cycle



## 7 AI Security Control Planes

| Identity, Delegation & Authorization | Data, Retrieval & Memory Integrity | Runtime Behaviour & Inference Assurance | Tool, MCP & Action Control | Execution Provenance & Forensics  | Resilience & Containment      | Governance, Supply Chain & Lifecycle  |
|--------------------------------------|------------------------------------|-----------------------------------------|----------------------------|-----------------------------------|-------------------------------|---------------------------------------|
| Who is acting, for whom?             | What can influence it?             | Is it behaving safely?                  | What can it do?            | Can we reconstruct what happened? | What limits the blast radius? | What can change, and who approves it? |

|                                         | Inception                            | Design                             | Build | Validate                                | Deploy                               | Operate & Monitor | Re-evaluate | Retire |
|-----------------------------------------|--------------------------------------|------------------------------------|-------|-----------------------------------------|--------------------------------------|-------------------|-------------|--------|
| Governance, Supply Chain & Lifecycle    | Governance, Supply Chain & Lifecycle |                                    |       |                                         |                                      |                   |             |        |
| Data, Retrieval & Memory Integrity      |                                      | Data, Retrieval & Memory Integrity |       |                                         |                                      |                   |             |        |
| Runtime Behaviour & Inference Assurance |                                      |                                    |       | Runtime Behaviour & Inference Assurance |                                      |                   |             |        |
| Identity, Delegation & Authorization    |                                      |                                    |       |                                         | Identity, Delegation & Authorization |                   |             |        |
| Tool, MCP & Action Control              |                                      |                                    |       |                                         | Tool, MCP & Action Control           |                   |             |        |
| Execution Provenance & Forensics        |                                      |                                    |       |                                         | Execution Provenance & Forensics     |                   |             |        |
| Resilience & Containment                |                                      |                                    |       |                                         | Resilience & Containment             |                   |             |        |

Leadership decides the boundaries. Architecture enforces them.

DO THIS MONDAY

# Five questions every CISO should ask

- 1 What can influence this AI?**  
Prompt, RAG, a webpage, tool output, another agent?
- 2 What does it remember?**  
Can untrusted information persist into later decisions?
- 3 Who is it acting for?**  
A human, a service, or another agent?
- 4 What can it actually do?**  
Read, write, send, delete — pay?
- 5 If it's wrong or manipulated, what limits the blast radius?**

# 30-60-90 Day AI Security Implementation View

30

**First 30 Days**  
Build visibility



- ✓ Inventory AI systems
- ✓ Classify data
- ✓ Classify authority
- ✓ Classify memory
- ✓ Classify tools
- ✓ Assess business impact

60

**60 Days**  
Reduce uncertainty



- ✓ Threat-model highest-risk systems
- ✓ Validate identity
- ✓ Validate delegation
- ✓ Check RAG / memory controls
- ✓ Verify action authorization
- ✓ Enable logging

90

**90 Days**  
Operationalize controls



- ✓ Introduce runtime monitoring
- ✓ Set execution budgets
- ✓ Prepare incident playbooks
- ✓ Define change / re-evaluation triggers



**Outcome:** from visibility → control validation → operational assurance

A practical 30-60-90 day view for CXOs, CISOs and security architects.

# Outcome

- Identify the AI systems with authority.
- Find the ones with the largest blast radius.
- Control what influences them.
- Make authority explicit.
- Gate consequential actions outside the model.
- Instrument the execution path.
- Re-evaluate whenever the system materially changes.

# Thank You



Download Resource

