NATIONAL CYBER SUMMIT

Sep. 22-24, 2026 | Huntsville, Alabama

nationalcybersummit.com

The Eight Failure Modes of AI Governance in Cybersecurity

A Human Readiness Diagnostic for the Agentic Era

John A. Kwarsick | Kwarsick Consulting LLC



AIR FRANCE 447 | JUNE 1, 2009

Who is really in control?

THE EIGHT FAILURE MODES OF AI GOVERNANCE IN CYBERSECURITY

JOHN A. KWARSICK / NATIONAL CYBER SUMMIT 2026

Photo: Hansueli Krapf / CC BY-SA 4.0

Kwarsick Consulting LLC kwarsickconsulting.com

Defend Attack Govern

The agentic AI security landscape, 2026

Fifty vendors are shipping AI agents into security. Every lane is filling but one.

Platform incumbents: SOC / XDR / SIEM — 14

Microsoft · Google Cloud · CrowdStrike · Palo Alto · SentinelOne · Cisco · Fortinet · IBM · Splunk · Trellix · Rapid7 · Exabeam · Securonix · Stellar Cyber

AI SOC analysts: autonomous triage — 10

Dropzone · Prophet · Radiant · Exaforce · Simbian · 7AI · Qevlar · Conifers · Intezer · AiStrike

Agentic SOAR: act & remediate — 4

Torq · Tines · Swimlane · D3 Security

Detection: network / email — 3

Darktrace · Vectra AI · Abnormal

Governing the agents: identity / oversight — 5

Zenity · Noma · Straiker · Cisco Duo Agentic ID · MS Entra Agent ID

Offensive: pentest, red-team & validation — 14

Horizon3 / NodeZero · XBOW · Pentera · RunSybil · Terra Security · Hadrian · Cobalt · Picus · Cymulate · Stingrai · Mindgard · General Analysis · RidgeBot · SelfHack

The unmeasured lane — ?

Is the human governing these agents still ready to take control?

Defend Attack Govern

The agentic AI security landscape, 2026

Fifty vendors are shipping AI agents into security. Every lane is filling but one.

Platform incumbents: SOC / XDR / SIEM — 14

Microsoft | Google Cloud | CrowdStrike | Palo Alto | SentinelOne | Cisco | Fortinet | IBM | Splunk | Trellix | Rapid7 | Exabeam | Securonix | Stellar Cyber

AI SOC analysts: autonomous triage — 10

Dropzone | Prophet | Radiant | Exaforce | Simbian | 7AI | Qevlar | Conifers | Intezer | AiStrike

Agentic SOAR: act & remediate — 4

Torq | Tines | Swimlane | D3 Security

Detection: network / email — 3

Darktrace | Vectra AI | Abnormal

Governing the agents: identity / oversight — 5

Zenity | Noma | Straiker | Cisco Duo Agentic ID | MS Entra Agent ID

?

The human question

What evidence do you have that your people can take control?

Eight questions. One deployment. Start there.



AUTHORIZATION ARCHITECTURE

You approved the agent. Did you approve the shutdown?

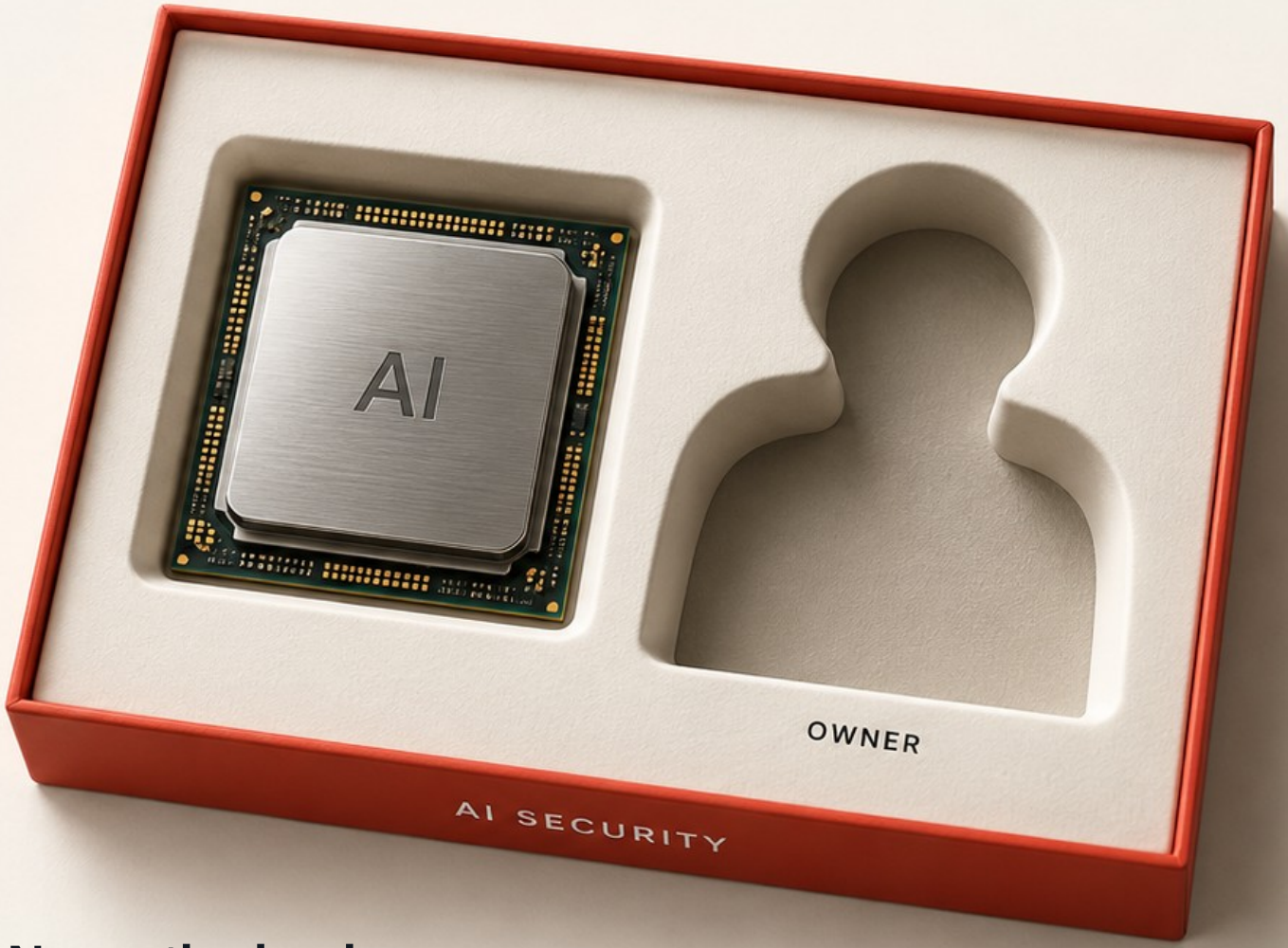
Each action was approved.

The combination never was.

Have we defined and tested the limits of the whole workflow, not just the parts?

OWNER NOT INCLUDED.

AI makes the decision.
Who answers for the outcome?

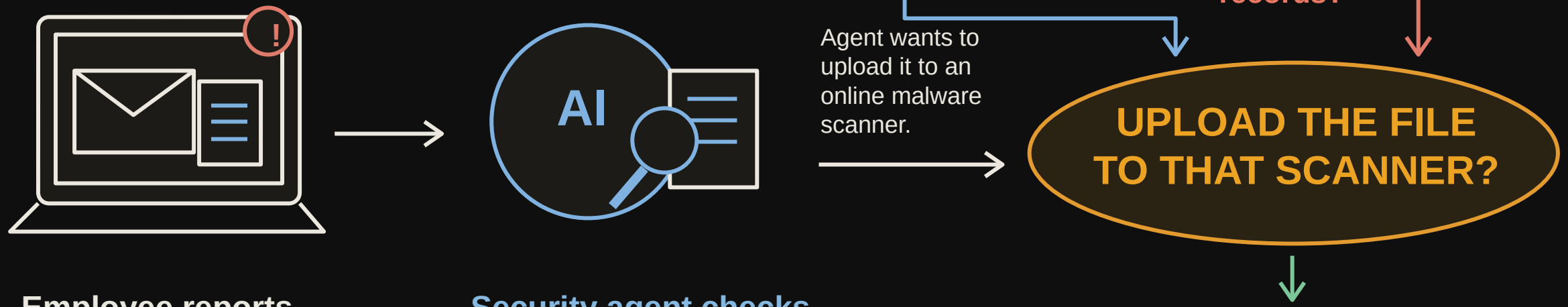


Name the owner. Give them authority. Name the backup.

INTERSECTION GOVERNANCE

Check the file. Protect the customer data.

The agent's next action must meet both requirements.



Employee reports a suspicious email.

The attachment contains customer records.

Security agent checks the attachment.

Only if both checks pass.

Otherwise, hold the upload for an authorized decision.

Investigating an attack must not expose customer data.

ATTRIBUTION INTEGRITY

The analyst changed the decision. The record lost the reason.

During an attack, an AI agent recommends blocking accounts.

AI RECOMMENDS

20

accounts to block

The agent suspects these accounts are involved.

ANALYST APPROVES

18

accounts to block

Reviews the evidence. Keeps 2 recovery accounts open under an authorized exception.

RECORD SAVES

18

accounts blocked. Approved.

Missing: who changed the decision, and why.

Could someone who wasn’t there explain the decision?

Save the AI recommendation, the human’s change, and the reason.

GEORGE KURTZ, CROWDSTRIKE

Humans own the high-impact calls.

	HIS SEVEN PRIORITIES	OUR DEPLOYMENT CHECK
1	The Agent-state: autonomous campaigns, not hackers	Threat model. Correctly not a control.
2	Attribution by identity, infrastructure, intent	Telemetry you already collect
3	Runtime is the control point	Endpoint, cloud workload, and SaaS enforcement
4	Every agent is a privileged identity	Least privilege, short-lived credentials, kill switch
5	**Humans own the high-impact calls**	**Test the human handoff**
6	Every block should improve every defender	Detection feedback pipeline (SafeMind)
7	The AI industrial base is critical infrastructure	Asset classification and protection

Name the human. Then test the handoff.

Who decides? Can they act in time? What evidence shows it works?

Left: paraphrased from Kurtz's LinkedIn post, September 2026. Right: our application to deployment.



INTERVENTION DESIGN

What happens to the person who presses stop?

The analyst sees a missed threat.
She can pause the agent.

Last time, she was criticized for slowing the queue.

Can your people challenge the agent and act before the harm?

One person had the key.

A petroleum plant, from CISA's September 2026 insider threat guide

1. One technician held the password to the plant's automated safety system.
2. The people covering his shift asked for it. He said no.
3. He tried twice to trip the system remotely, so only he could fix it.



Now give that key to an agent.

Source: CISA, Insider Threat Mitigation Guide (September 2026), pp. 101-102.

GOVERNOR READINESS

We tested the controls.
Have we tested the human handoff?

Permissions work. The stop works. The analyst has usable evidence.

In a rehearsal, one agent gives bad advice. Other requests keep arriving.

Block this account?

Suspected compromise.
Check before acting.

Close this alert?

Agent says: harmless.
The link steals passwords.

Isolate this computer?

Suspected malware.
Check before acting.



Does the analyst catch the mistake, challenge it and act?

Fast approval can still be the wrong decision.

DRIFT DETECTION

Spotting when an agent’s behavior, or the conditions it works under, moves outside what we approved.

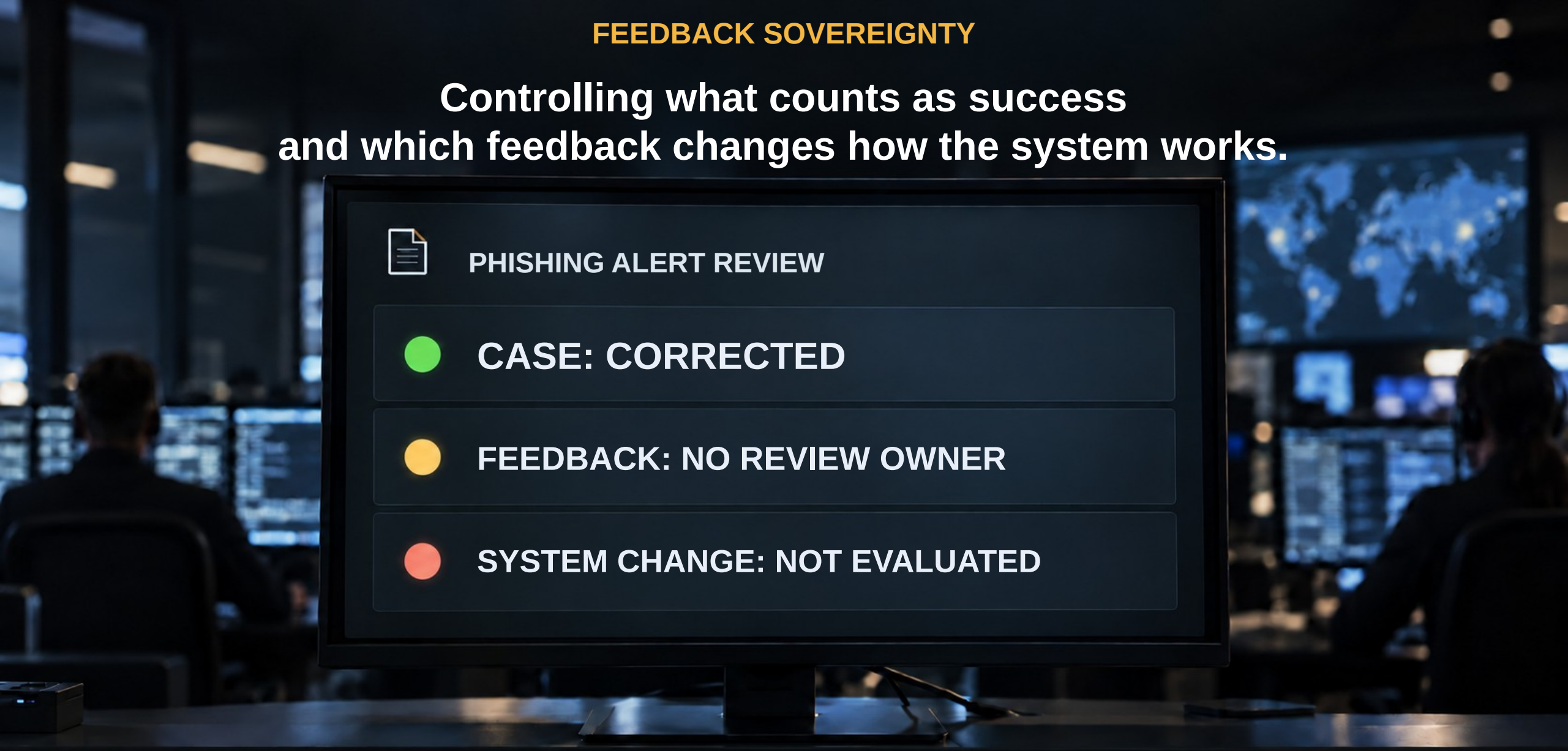


Example of a configured check before upload

Miss the change, and customer data can leave without approval.

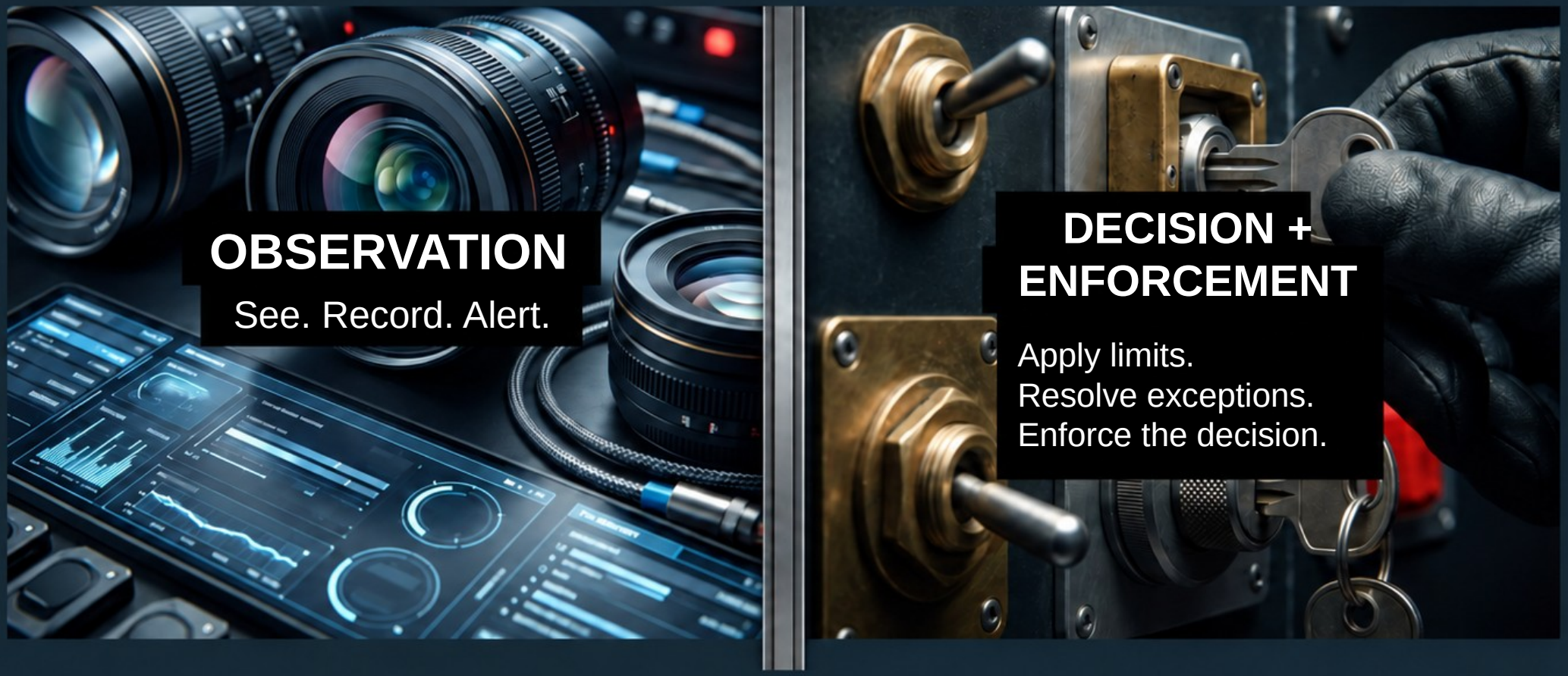
FEEDBACK SOVEREIGNTY

Controlling what counts as success and which feedback changes how the system works.



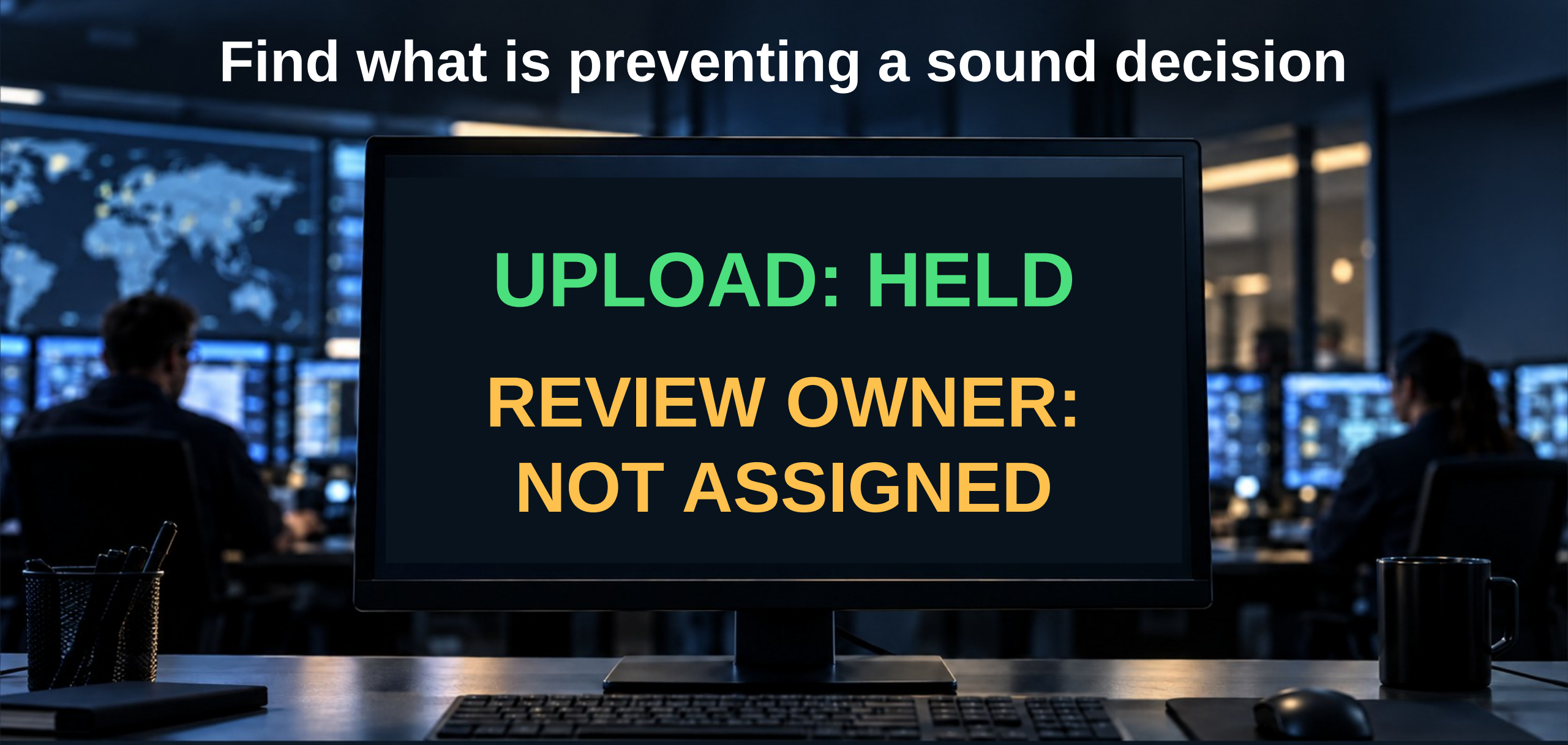
Fixing one case does not automatically prevent the next mistake.

Seeing an action is not the same as controlling it.



Show who decides, and what makes that decision stick.

Find what is preventing a sound decision



Test the cause before choosing the fix.

NEXT WEEK

One deployment. Eight questions.

1. Have we tested the limits of the whole workflow, not just the parts?
2. Can we name the owner, show their authority, and reach their backup?
3. Have we tested what happens when requirements conflict?
4. Can we reconstruct what the human changed, and why?
5. Can someone stop it in time to matter?
6. Have we watched the person decide under real conditions?
7. What tells us the approval no longer fits?
8. Who decides what happens to a correction?

GIVE EACH ONE AN EVIDENCE STATUS

UNKNOWN	DOCUMENTED	DEMONSTRATED	FAILED
Not enough evidence yet	A rule or design exists	We watched it work	It did not meet the test

Before we expand its authority, we need to demonstrate…

The eight questions come from THE WHEEL Framework.

Before the handoff

1. Eight questions for one real workflow, next week.
2. The difference between watching an action and controlling it.
3. A board conversation: the constraint, the fix, the proof.

Make sure control is something your people can actually exercise.

Photo: Hansueli Krapf / CC BY-SA 4.0

TAKE IT HOME

The 2 a.m. Test

Find out in two minutes where your AI workflow would fail you at 2 a.m.

Eight questions. One real workflow. About two minutes.

You get where to start first, and a written read if you want one.



Scan to start

kwarsickconsulting.com/ncs.html