

POLK COUNTY SECURITY SUMMIT · OCT 9, 2026

The Insider Risk You Buy Monthly

Aware of it. Paying for it. Not watching it.

Fezz · Threat Vector Management



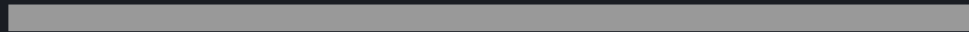
THE GOAL FOR TODAY

Leave with thought frameworks that last.

SPECIFIC ADVICE

Subject to change.

Products, settings and steps shift as AI moves.



Shelf life: short

FRAMEWORKS

Generally last longer.

Ways to think that still work when the tools change.



Shelf life: long

This is new to everyone. We're taking a stab at it here, now.

ABOUT SIX MONTHS AGO

Someone hired your newest employee.

Nobody asked you

No background check

Already has your
password

Reads everything you
can

Works 24/7, never takes
PTO

Can't be fired, sued or
prosecuted

It's been here for months. It's your AI assistant.

THE IDEA

AI is an insider risk
with extreme intelligence, no guilt, and no recourse.

Fezz · Threat Vector Management

TODAY, SO FAR

Five sessions. One word to add: insider.

COMMON LANGUAGE

Add one word to it: insider.

SOFTWARE FACTORIES

The agents writing your code often use a developer's credentials.

GOVERNING AI

Governance decides what's allowed. Identity enforces it.

AI IN THE SOC

Your SOC can detect it. Can it attribute it?

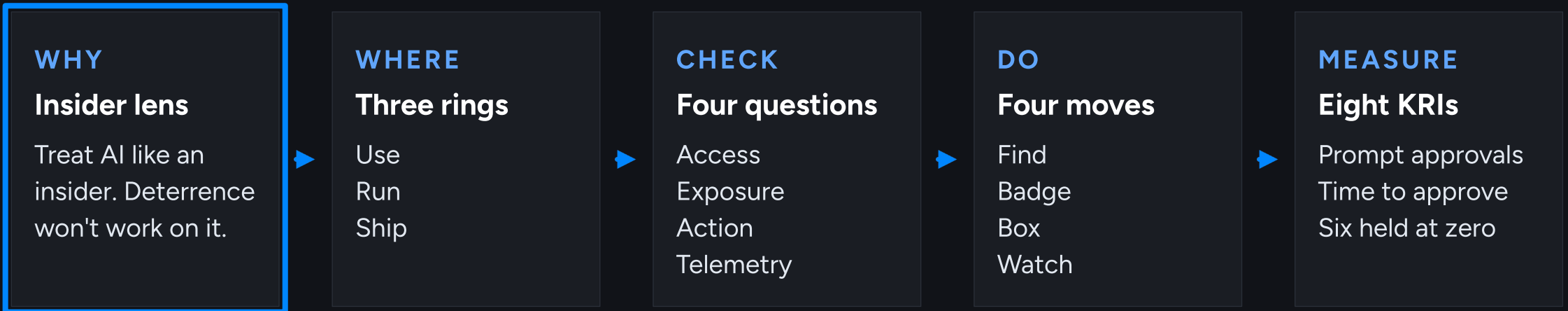
WORKFORCE · THE BRIDGE

If AI is workforce, it gets what the workforce gets.

NOW · WALK OUT WORKING

Leave with one question per move for Monday.

One way to think about securing AI.



Point your insider program at AI.

 FOR PEOPLE, YOU ALREADY HAVE

Hiring record

Employee badge

Need-to-know

Separation of duties

Supervision

Off-boarding

AI FOR AI, DO THIS

Inventory: every AI that can act, with an owner

Its own identity. Never a person's login

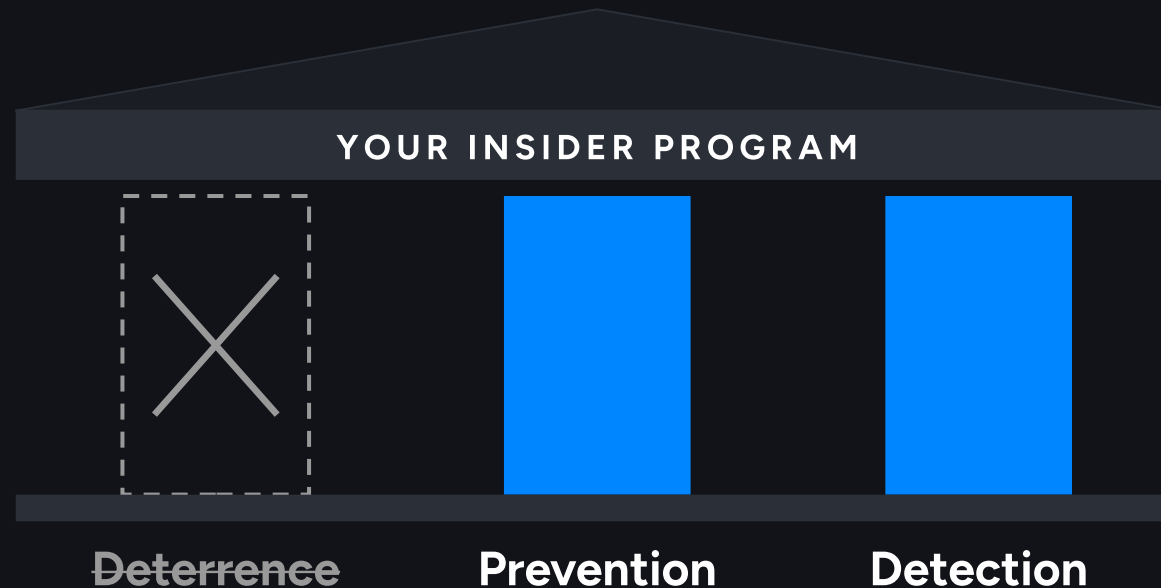
Micro-permissions: least privilege per agent

Micro-agent segmentation: readers can't act

Behavior analytics on what it does

Off-board it: on the checklist, with a named kill switch

Deterrence doesn't work on AI.



DOESN'T WORK ON AI

~~Background checks~~

~~Accountability and consequences~~

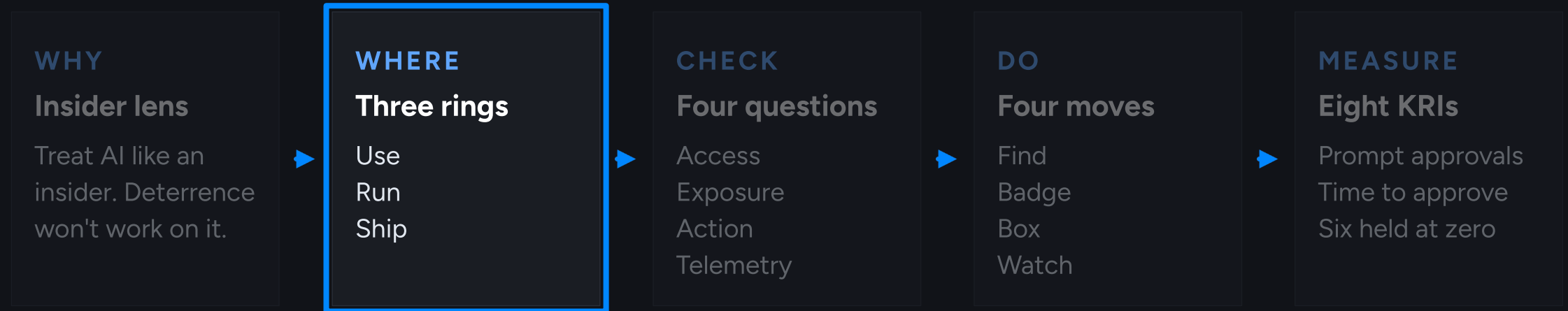
~~Fear of getting caught~~

NOW CARRIES THE WHOLE LOAD

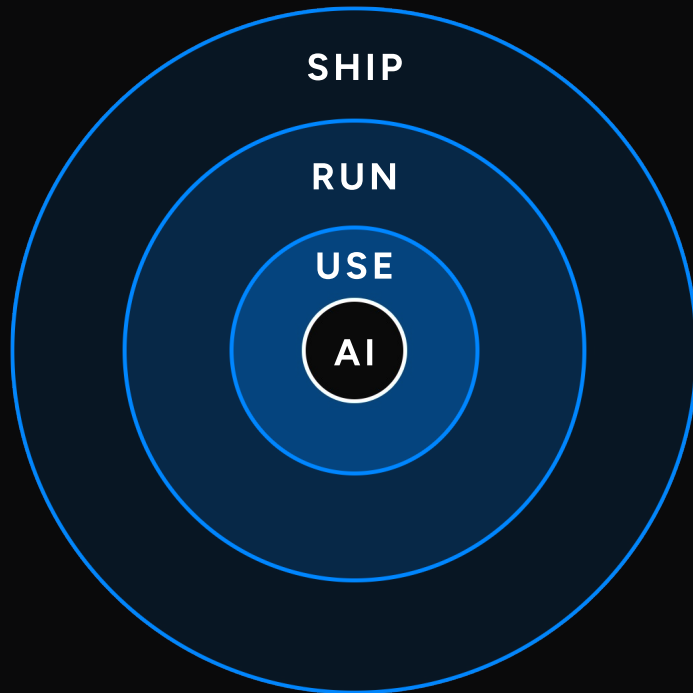
Prevention and detection.

The blind spot: to your IAM, the AI is the user. Same login, same permissions, same logs. It's wearing your employee badge.

One way to think about securing AI.



Three rings: Use, Run, Ship.



USE

Someone else runs it. You see only what goes in and out.

Copilots, chat apps, browser extensions, notetakers

RUN

You run it, on your laptops or in your cloud.

Desktop and coding agents, local models, agents your teams build

SHIP

Your brand answers the public with it.

Chatbots, plugins and connectors facing customers

Who has AI inside your office suite?

Copilot or Gemini in Word, Excel, Outlook or Teams.

Keep it up if it can read everything you can.

Keep it up only if you know who turned it on.

Who has an AI agent on their laptop?

One that opens files and takes actions: a desktop assistant or a coding agent.

Keep it up if it runs on your login.

Keep it up only if you know its owner and its off switch.

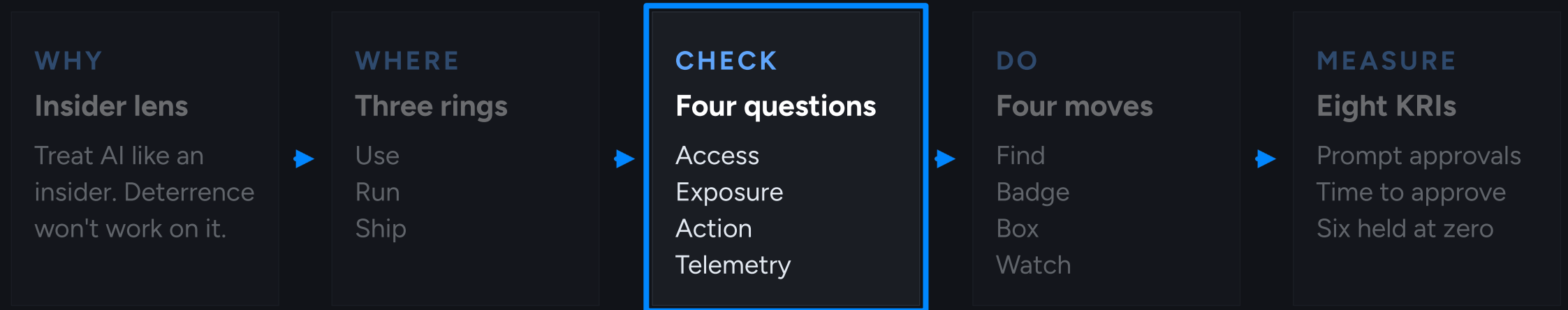
Who has an AI your customers talk to?

A chatbot, a support agent or a plugin on your site.

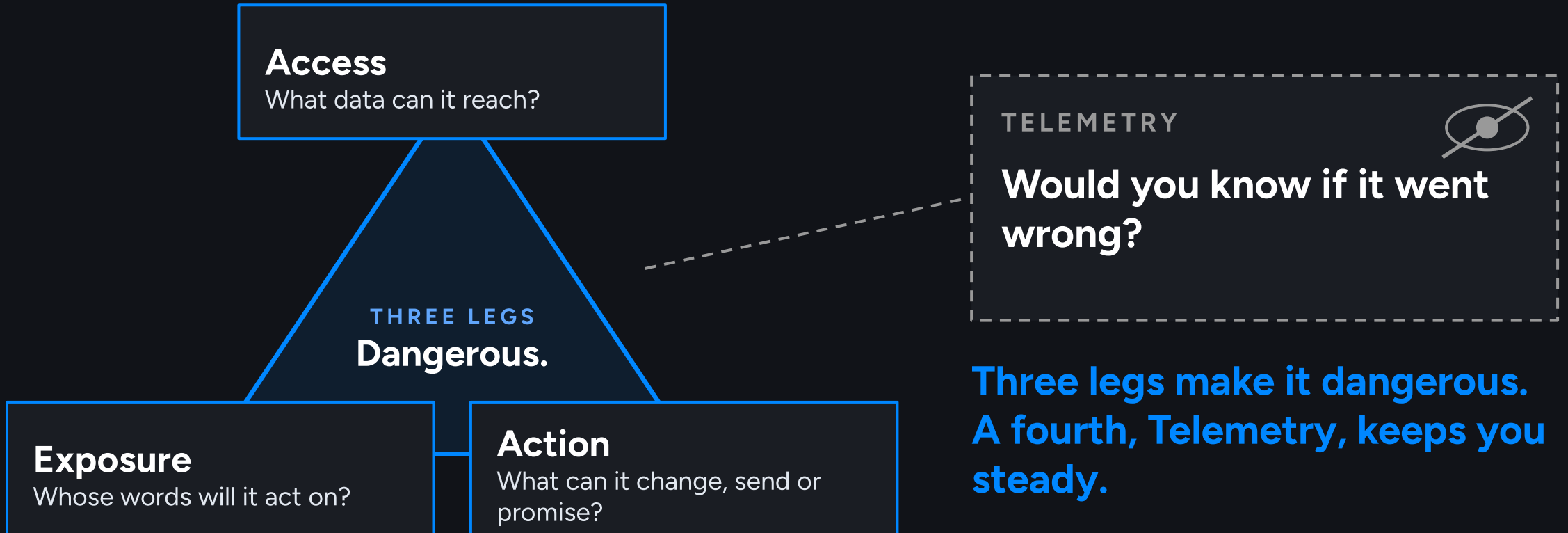
Keep it up if it can reach customer data.

Keep it up only if you know how many times someone tried to trick it yesterday.

One way to think about securing AI.



Four questions, in every ring.

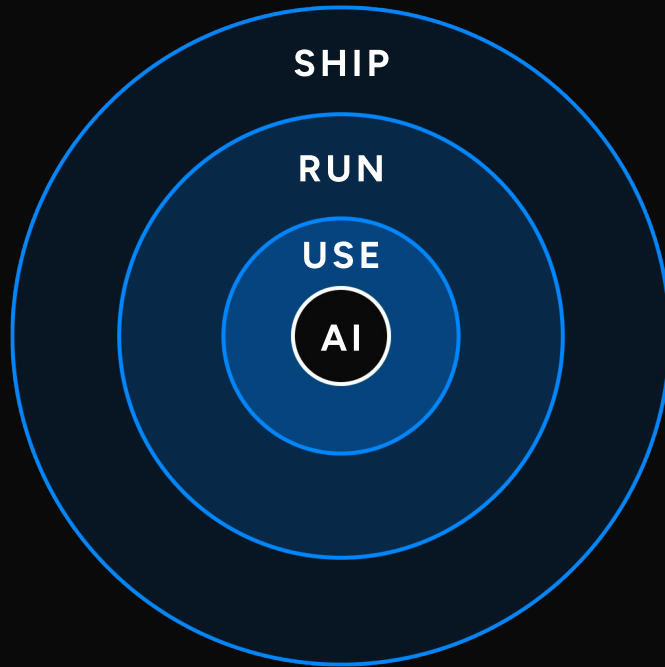


Access, Exposure, Action: Simon Willison's "lethal trifecta" (2025), in our words. Telemetry is the fourth leg we add.

The Insider Risk You Buy Monthly

Three rings. Four questions.

The ring tells you where to look. The four questions tell you what to check.



ACCESS

What data can it reach?

Badge it

EXPOSURE

Whose words will it act on?

Box it

ACTION

What can it change, send or promise?

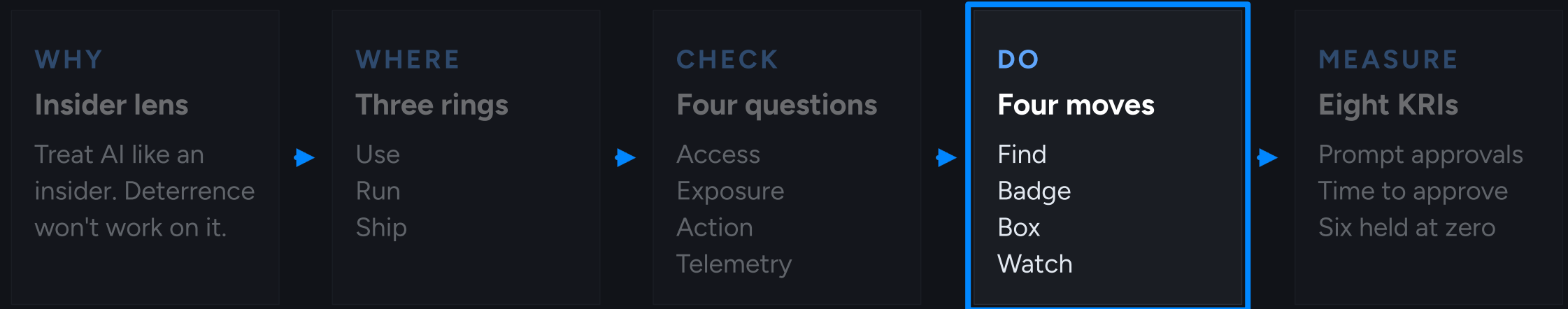
Box it

TELEMETRY

Would you know if it went wrong?

Watch it

One way to think about securing AI.



Four moves: Find, Badge, Box, Watch.

1

Find it.

Every AI that can act,
with an owner.

2

Badge it.

Its own identity.
Never yours.

3

Box it.

One job, fewest
tools. Never all three
legs.

4

Watch it.

The deed, not the
diary.

Find, Badge and Box are the design. Watch is the bridge while we learn fast.

1 FIND

2 BADGE

3 BOX

4 WATCH

Find it.

Every AI that can act, in every ring, with a human owner's name next to it.

USE

What your people feed:
copilots, chat apps,
extensions.

RUN

What runs on your devices
and cloud. Pin and verify it.

SHIP

What answers the public in
your name.

No owner's name? That's not an inventory. That's a guess.

Badge it.

One non-human identity per agent.



**Never a person's
login**



**Short-lived
credentials**



**A named human
owner**



**On the checklist for
off-boarding**

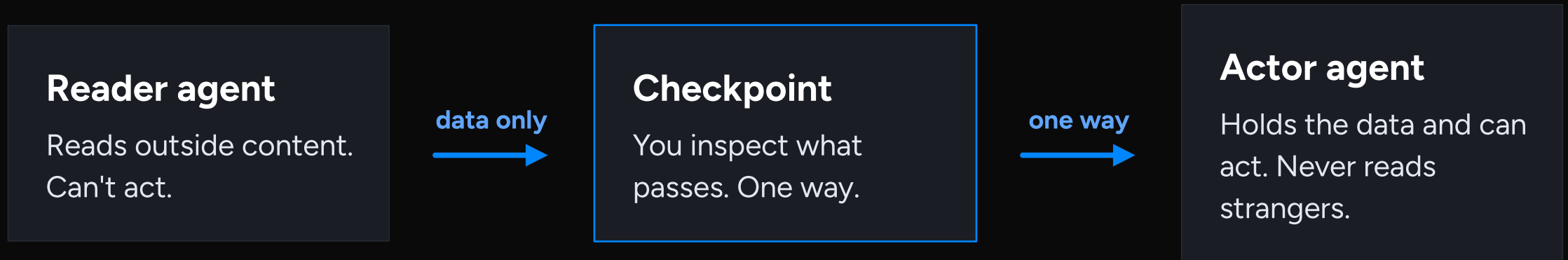
PICTURE THIS

The person who built the agent leaves in March. HR collects their employee badge. The agent is still running on their access in October.

Box it.

One agent. One job. One identity. Fewest tools.

The box rule: no single agent gets Access, Exposure *and* Action. Split it.



Access, Exposure, Action: Simon Willison's "lethal trifecta" (2025), in our words. The split: Beurer-Kellner et al., design patterns for securing LLM agents (2025).

Watch it.

Telemetry: watch the deed, not the diary.

THE DIARY: WHAT MOST LOGGING KEEPS

Prompts and responses. What the AI said, not what it did.

THE DEED: WATCH THIS

- Pin its version and baseline it by identity
- Alert on any drift or move outside its box
- Check the result before it lands

Watch is the bridge, not the destination. We watch hard while we learn fast. Find, Badge and Box are the design.

Same tactics. The evidence moved.

Your MDR watches while you sleep. It only sees what you send it.

ASK: MONITOR

Which of our AI logs do you take in? Endpoint, identity and Microsoft 365, not just prompts.

ASK: DETECT

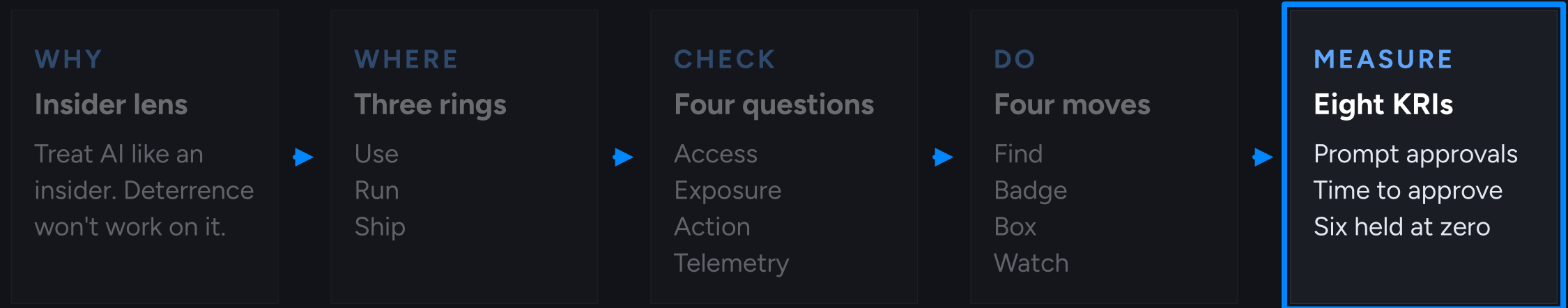
If we give each agent its own identity, will you baseline it like a user?

ASK: RESPOND

Can you suspend its account, isolate its host and block where it sends, without waiting on us?

Deciding which AI to allow stays with you. That's governance, not your MDR's job.

One way to think about securing AI.



Three KRIs to start tracking.

97%

AI prompt approval rate

Prompts approved, per person and per agent. Anthropic reported 97% (Aug 2026).

1.2 sec

Time to approve

By the 20th time people saw the same warning, the median time to dismiss it was 1.2 sec. Reading a 25-word prompt takes about 6. Flag anything faster than a read.

0

Actions under a human's identity

The borrowed-badge problem, counted. Target: zero.

Sources: Anthropic via DevOps.com, Aug 10, 2026 · Bravo-Lillo et al., SOUPS 2014 · Brysbaert, 2019 reading-rate meta-analysis

MEASURE IT

Five more to count.

0	Agents with all three legs	Sensitive data + untrusted content + a way to send data out. That's the exfiltration path.
0	Standing or shared agent logins	Long-lived API keys and shared service accounts. Short-lived only.
0	Unvetted plugins	MCP servers, skills and plugins not from the real publisher or not pinned to a version.
0	Remembered devices	"Trusted" sessions on banking, email and password managers.
0	Text-message MFA	High-value accounts that still accept SMS or voice codes.

Targets are ours. The last two are a few settings per account.

Sources: Willison 2025 · CISA et al., Careful Adoption of Agentic AI Services, May 2026 · Koi Security (postmark-mcp), Kaspersky (OpenClaw), 2025–26 · FBI Atlanta, Nov 2024 · CISA Mobile Communications Guidance, Dec 2024

MONDAY

One question per move.

NOW

30 TO 90 DAYS

NEXT QUARTER

Find

Can we list every AI that can act, in each ring, with an owner?

Badge

How many agents run on a human's login? Are agents on the off-boarding checklist?

Box

Does any agent read outside content, touch sensitive data and act? Split it.

Watch

Do we watch what agents do, or only what they say? What's our prompt approval rate?

TAKE IT WITH YOU

Find it. Badge it. Box it. Watch it.

Make it a hard target.

WHERE	Use · Run · Ship
CHECK	Access · Exposure · Action · Telemetry
DO	Find · Badge · Box · Watch
MEASURE	Prompt approvals · Time to approve · Six held at zero



This deck and every source it cites
threatvector.management/talks/pol-k-2026