



AI AGENTS IN GRC & CYBERSECURITY

Safe by Design

Building and governing AI agents that deliver.

Luke Reynolds · Product, Decision Focus

August 2026

Your people are already using AI. You just can't see it.

At a conference, risk and audit professionals told me — unprompted — how they really work:

1

They export your data

Pull risk and control data out of the official system, by hand.

2

Paste it into ChatGPT

Ask a public, consumer tool to transform it — then paste the result back in.

3

Against the rules

Some worked under an anti-AI policy. Others were limited to IT-approved tools — and did it anyway.

They weren't reckless — they had no sanctioned way to use AI. And when AI use is external, you cannot govern it: no boundaries, no context controls, no audit trail. The policy didn't stop it. It just pushed it into the dark.

THE TURN

A safe in-house agent fixes both halves of that problem

Give people a sanctioned way to do the work, and the shadow path disappears — because the official one is now faster.

Faster

No export, no copy-paste, no re-import. The agent works where the data already lives — so the sanctioned path beats the workaround.

Safer

Built deliberately: the right boundaries, the right context, and an audit log of everything it did. Governed by design, not by policy alone.

You can't police AI out of existence. You can give it a safe front door — and that's a design problem.

AND THEY'RE ALREADY HERE

Real work, real time back — in supplier onboarding

My examples come from onboarding ICT service providers: vetting and approving the third parties your business depends on.

70%

Faster onboarding

On cases that don't need full human review. Internal estimate, prototype data.

The point of an agent is the hours it gives back
— not novelty.



Wasted capacity

Skilled people buried in admin instead of high-stakes risk work.



Hidden cost

Expensive expert hours spent on low-risk, low-impact suppliers.



Opportunity cost

Your most critical reviews delayed by a backlog of routine ones.

Accelerate the experts. Don't replace them.

People stay accountable for the decision. Agents do the grind — and everything they do is attributable.

1

Humans own the call

Instruct the agent on its do's and don'ts, like an employee. You decide which decisions it may make, and how it makes them.

2

Agents own the grind

Reading questionnaires, summarising findings, scoring risk, drafting amendments — all the repeatable lift.

3

Everything is attributable

Each agent runs under its own identity, not a person's login. Every action is logged with its reasoning and the exact context it used.

Four things to get right

A lifecycle, not a checklist: justify it, equip it, constrain it, watch it. We'll carry one agent through all four.

1

Why

Define the problem and the value. Weigh reward against risk. No agent for its own sake.

2

Context

Give it the situational knowledge — and the instructions — to reason well.

3

Boundaries

Decide what it can do vs only surface. Keep a human gate where it matters.

4

Monitoring

Log everything — for audit and for value. Spot drift; prove what's working.

Human review threads through steps 3 and 4 — and becomes the whole of the governance story.

Start with the question most teams skip

Every governance tool captures risk in detail. Few ask what the agent is for, or what value it will bring.

Deciding on half the picture

- A high-risk agent that meets three strategic goals gets rejected — because no one asked about the upside.
- A low-risk agent with no real value gets approved by default — and quietly adds risk for nothing.
- **The bad trade is high autonomy or access with little reward to justify it.**

The rule

Value proportional to risk.

Capture why you want the agent and the value it should deliver at the same moment you assess its risk — and put both on one screen.

Time saved is the most concrete form that value takes — and the easiest to defend later.

Brief it like a new hire — no code, no retraining

You wouldn't ask a junior to review your risks and controls without telling them how. Neither should an agent run blind.

1

Who you are

Your industry, markets, regulators, customer base and risk appetite — read by the agent on every run.

2

Your rules

Your tiering policy, your due-diligence rubric, your legal checklist. It reasons from your rules, not ours.

3

This request

Why this supplier, what service is being bought, and which part of the business will use it.

We make proper instructions a prerequisite for every agent — plain-English, versioned, customisable per client. Edit it, re-push: the agent's reasoning updates with no retraining.

Decide what it can do — and what it can only surface

The safest pattern: the agent recommends, the human decides where it matters. Default to surfacing when stakes are high.

SURFACE FOR REVIEW

TAKE ACTION



Our reviewer agent lives here — it drafts findings, and never auto-advances onboarding or replies to the supplier.

Aggression

What gets auto-progressed vs flagged. e.g. progress only when impact × likelihood is low enough — and you set that threshold.

Scope

Which requests it even picks up — using filters. Exclude tiers, regions, supplier types, contract values.

Human gates

Where a person must sign off. On rejections only? On every finding? On every contract above a set value? Your call.

Nothing happens off the record

Every run logged — a safety record and a value record at once. This is the hinge into governance.

AGENT RUN LOG

Agent	Due Diligence Reviewer
Decision	Findings raised → human queue
Model	gpt-4o
Instructions	Client review rubric v1.3
Time saved	42 min
Human review	Approved — 2 of 3 findings kept
Status	Success



The exact instructions used

Every prompt and piece of context sent to the model is captured — fully reproducible.



Time saved, added up

Minutes per run, summed against your hourly cost. Value, evidenced.



The human verdict

The review is captured — and it's what sharpens the agent over time.

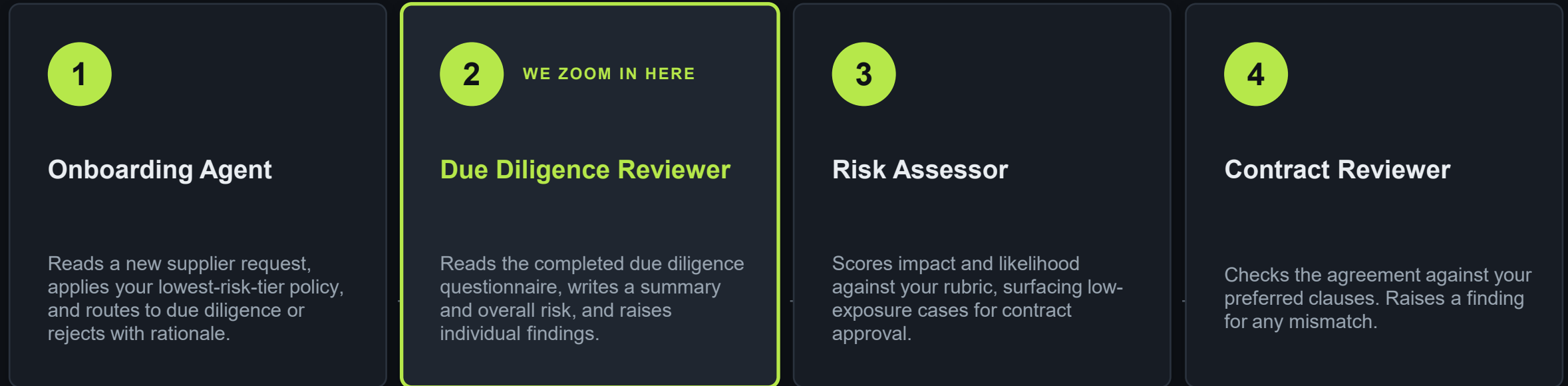


A link to the run

One click from the audit row to the underlying execution.

Four agents, one journey — and not one triggers the next

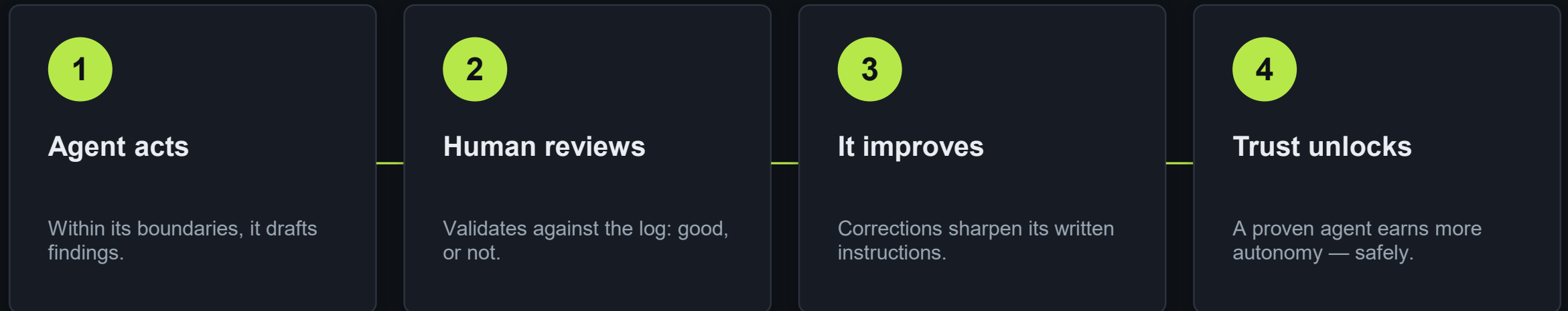
Each handles part of the onboarding lifecycle. Every handoff needs a human before the next begins.



Segregation of duties, applied to agents. As clients come to trust the accuracy over time, those human gates can open — and more of the chain can run automatically.

The human review is where trust is earned

The agent acts within bounds, produces a draft, and a human validates it. That validation is data.



The loop closes: Step 1 asked "is it worth it?" — by Step 4 you have a measured answer. Trust isn't asserted; it's accumulated, one reviewed decision at a time.

One pane of glass for your AI workforce

Aggregate the logs and the question changes from "did this run work?" to "which agents have earned trust?"

1,284

Runs this month

312h

Time saved

94%

Auto-progress rate

0.4%

Failure rate

This is the bridge. An agent's measured accuracy is exactly what a decision-maker needs to govern a growing estate: which to widen, which to rein in, which to retire. But it raises a harder question —

if the human review isn't good... who reviews the reviewer?

What's broken in AI governance today

The same friction that pushes one person to paste data into ChatGPT scales to the whole organisation.

Approval bottlenecks

Slow, fragmented processes drive users to bypass governance entirely.

Shadow AI

People use public GenAI because the official path is too slow — exactly what we opened with.

Risk without reward

Tools capture risk in detail; few ask what the AI is for. Approvers see half the picture.

No 2nd-line validation

Regulators expect independent challenge. Today it's ad-hoc spreadsheets.

Duplicate effort

Three teams build the same use case. No shared inventory.

Fragmented accountability

No single owner across Legal, Risk and IT. Stalled decisions, control gaps.

Governance asks the same four questions

What you ask when designing one safe agent is what the organisation should ask of every AI it runs.

DESIGNING ONE AGENT

- 1 Why — what value, at what risk?
- 2 Context — does it have what it needs?
- 3 Boundaries — what may it do alone?
- 4 Monitoring — is it working, provably?



GOVERNING THE ESTATE

- 1 Intake captures why & projected value
- 2 Born linked to its data, vendor, service
- 3 Approval sets autonomy & human gates
- 4 Independent validation + value realised

Second-line validation: independent challenge

First-line review by the business is necessary but not sufficient. An independent validator provides the assurance.



Validator ≠ Owner

Enforced on the record — the named validator cannot be the named owner.



Cadence by materiality

Higher-impact systems are revalidated more often. The schedule is set by risk.



Structured evidence

Soundness, benchmarking, outcomes, limitations, drift — captured the same way every time.



Linked to approval

Findings link back to the approval that started it. The trail stays continuous.

One validation record, evidence multiple regulators recognise — the EU AI Act and ISO 42001 both expect independent challenge, not the use-case owner marking their own homework.

The Projected → Realised value loop

Forecast at intake, measured at cadence, recorded with audit-grade evidence — the same value question, run for the whole estate.

Projected

Captured at intake — baseline, target, timeframe.

Approved

Risk and reward weighed on the same screen.

Measured

Realisation records at cadence: actual vs forecast.

Defended

Achieved / Exceeded / Partially / Missed — with evidence.

"My annual bonus is now tied to how much measurable value we get from AI in our day-to-day processes."

— Chief Risk Officer, large global insurer

THE TAKEAWAY

Seven principles for a safe agent

If you photograph one slide, make it this one — in the order you'd actually think them through.

1

Value proportional to risk

Don't grant autonomy or access an agent hasn't earned in time saved.

2

Brief it properly

You wouldn't hand a junior your risk review without telling them how. Make instructions a prerequisite.

3

Default to surfacing

Where stakes are high, the agent recommends — it doesn't act.

4

Computers aren't accountable

It may suggest or even enact a decision — only if it's gated, reviewable, and triggers nothing downstream.

5

No agent triggers another

A human reviews every handoff in the chain.

6

Review to refine

The human verdict is a signal — logged corrections sharpen its written instructions.

7

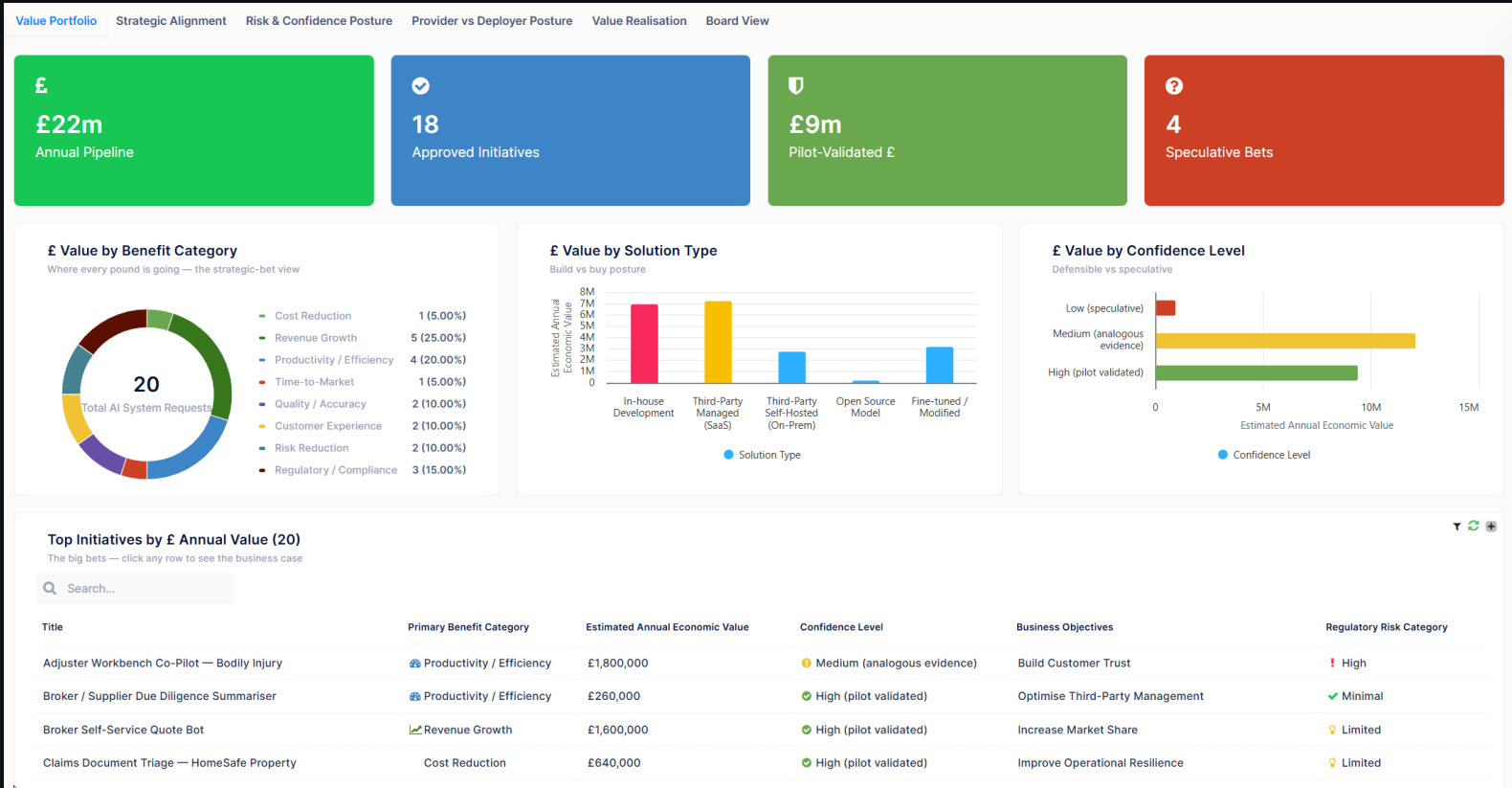
Earned autonomy

Start constrained. Proven accuracy is what loosens the reins, over time.

THE QUESTION EVERY BOARD NOW ASKS

Is the AI actually worth it?

Governance doesn't start with risk. It starts with the same question you ask when designing one safe agent: why, and what value? Here's the very first screen of an AI request



Ask it before you build. Bake it into the first "can I use this?" — and risk-versus-reward becomes culture, not paperwork.