We Hold Agents to a Higher Standard Than Ourselves. Good.

Autonomy Platform · Public Blog

CONFIDENTIAL

Industry June 2026

We Hold Agents to a Higher Standard Than Ourselves. Good.

Ram M catalogues five ways good employees accidentally weaken their organizations, each with a hidden cost. We hold AI agents to a higher standard than we hold ourselves, and there are four reasons that is exactly right.

This week Ram M published a clear-eyed piece, *How Employees Accidentally Cripple Organizations*. Its argument is careful and humane. Companies are rarely brought down by a few bad people or one catastrophic decision. They are worn down by thousands of small, reasonable actions taken by good employees trying to do their jobs. The word that carries the whole piece is *accidentally*. Nobody intends the harm. It emerges from incentives and structure and the ordinary wish to avoid risk and protect a relationship.

He names five mechanisms: self-preservation, local optimization, information distortion, dependency creation, and risk avoidance. Against each one he sets a column he calls the *Hidden Cost*: the project slip that

doubles before anyone admits it, the departmental win that becomes an enterprise loss, the decision leadership makes on an incomplete picture, the knowledge that walks out the door, the opportunity that passes while everyone waits for certainty. The harm is real and quantifiable. It is just never invoiced to anyone.

I read the list a second time and noticed something. Every one of those five behaviours is something we have already decided is unacceptable in an AI agent. We write requirements against them. We refuse to deploy systems that do them. We hold agents to a standard of conduct that the organisations deploying them have never been held to.

We hold agents to a higher standard than we hold ourselves.

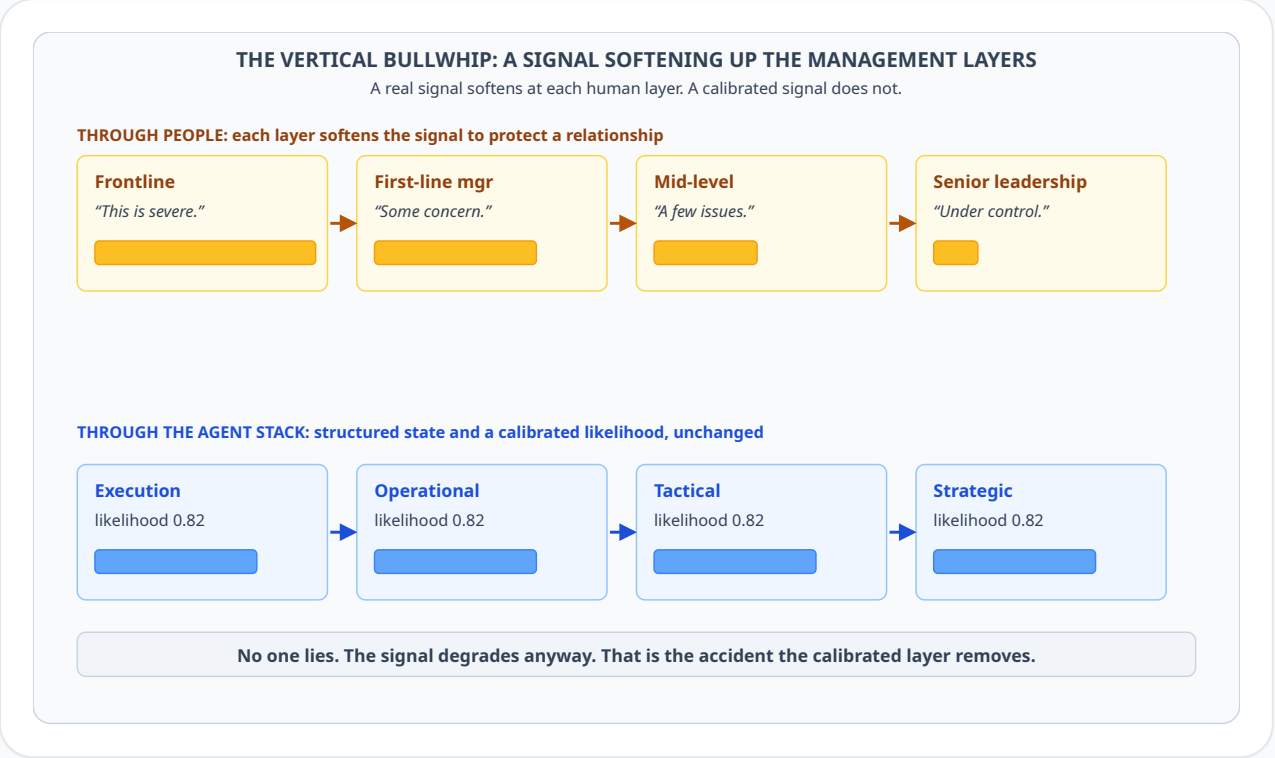
> **We hold agents to a higher standard than we hold ourselves. That is not a double standard to apologise for.**
>
> It is the right standard, and keeping it is the entire point.

This post is about why that is a good thing, and worth defending.

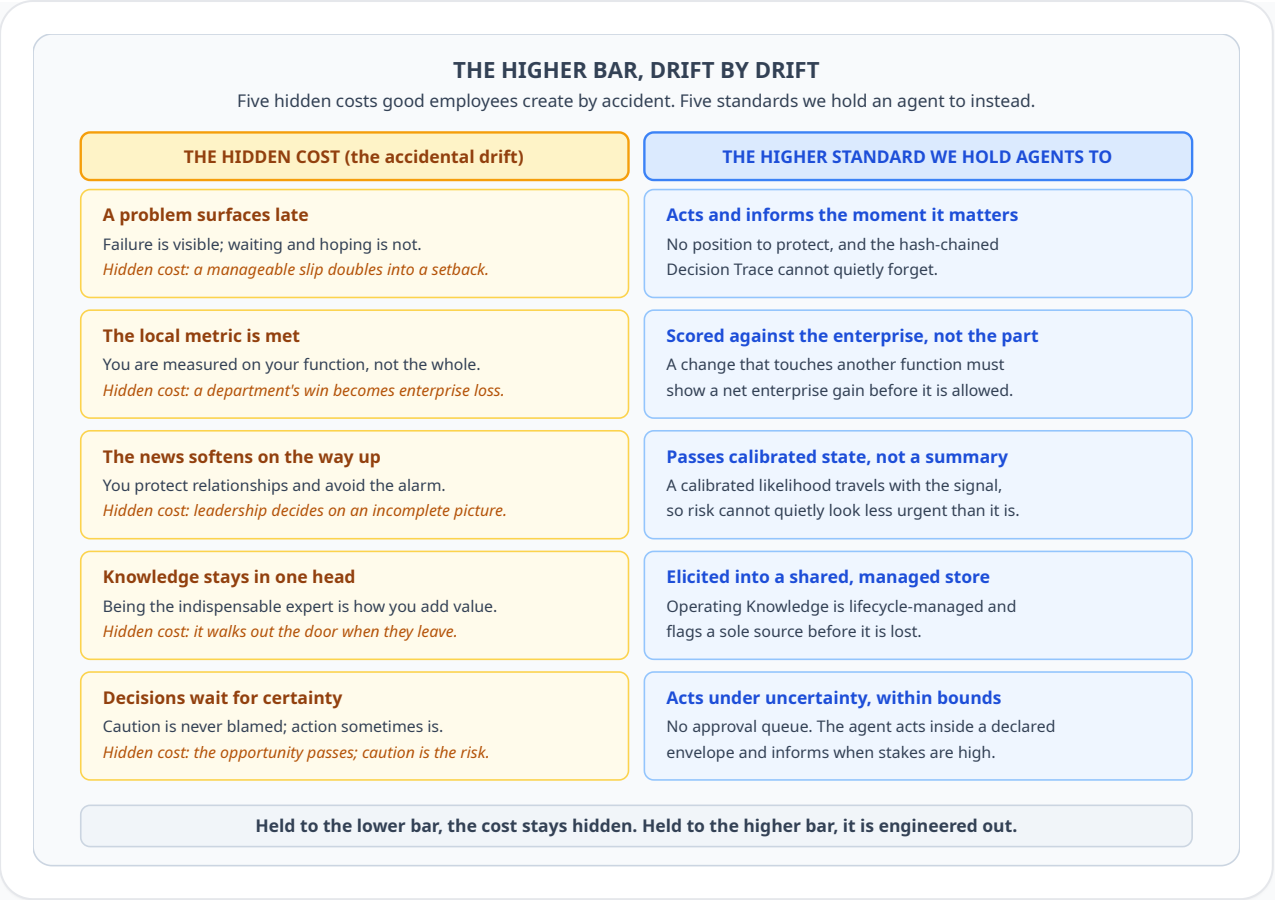
The higher bar, drift by drift

The supply-chain reader will recognise the third mechanism, information distortion, immediately, though it is worth being precise about which bullwhip we mean. The classic bullwhip effect is *horizontal*: demand

variance amplifying upstream across the echelons of a supply chain, because each link optimises locally without seeing true end demand. What Ram M describes is its *vertical* cousin: the signal softening as it climbs the management layers, until leadership sees a reassuring version of reality. Same mechanism, a different axis. The diagram below is the vertical form; I come back to the horizontal one, and what Autonomy does about it, further down. No one lies. The signal degrades anyway, and the Hidden Cost lands on whoever decides at the top.



That is the tell for all five. They are not failures of character. They are what happens when you route decisions through people who are, quite reasonably, protecting a position, a metric, a relationship, or a reputation. So take each in turn, with no blame attached to anyone. On the left, the accidental drift and the hidden cost Ram names. On the right, the standard we hold an agent to instead.



Note what the right column is not. It is not “agents are honest and people are not.” Each standard is something we insisted on when we built the agent, and never insisted on when we built the organisation. The agent does not surface the bad news because it is virtuous. It surfaces it because it has no career to protect and no review meeting to survive, and because the Decision Trace it writes to is append-only. We engineered the absence of the pressure that makes a good person wait.

The other bullwhip, and what radical visibility does to it

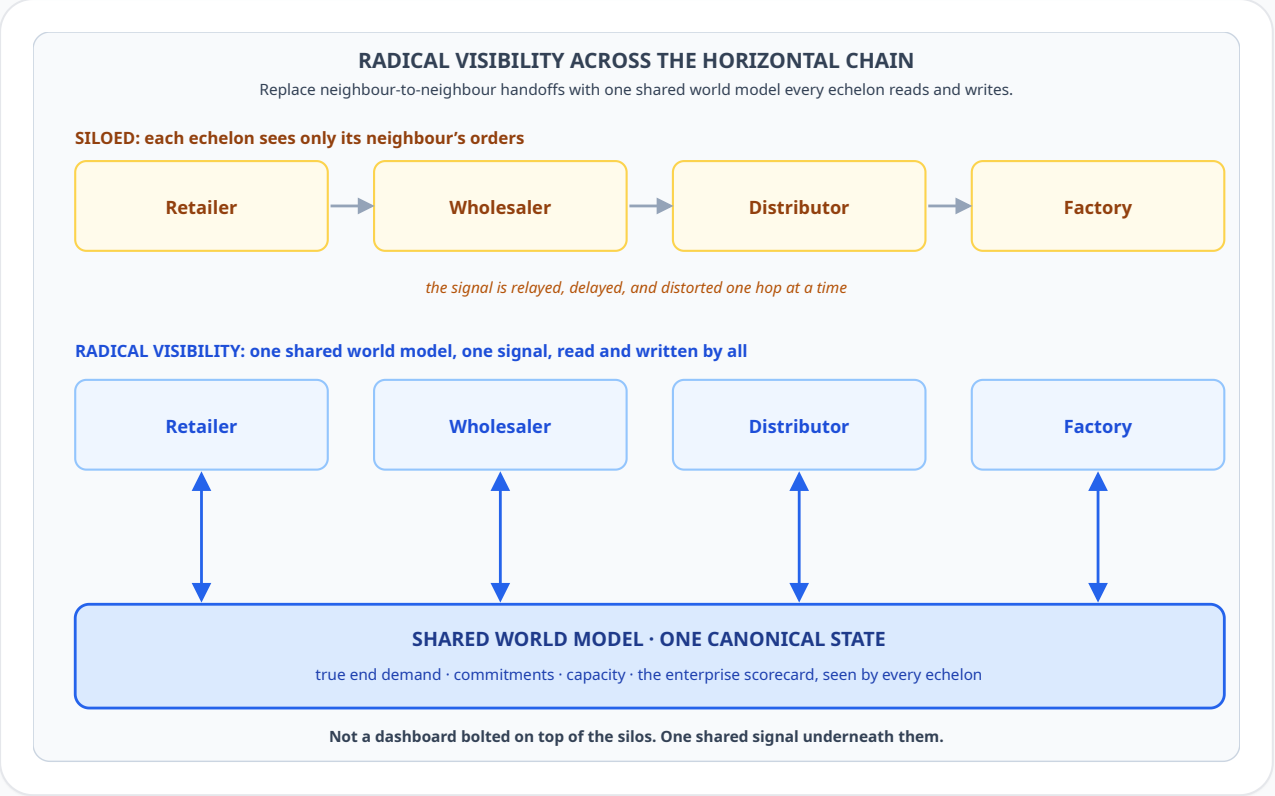
The vertical bullwhip is the organisational one. The *horizontal* bullwhip is the one supply-chain people have studied for sixty years, and it is the one Autonomy was built to damp. Jay Forrester identified it in 1961 and created

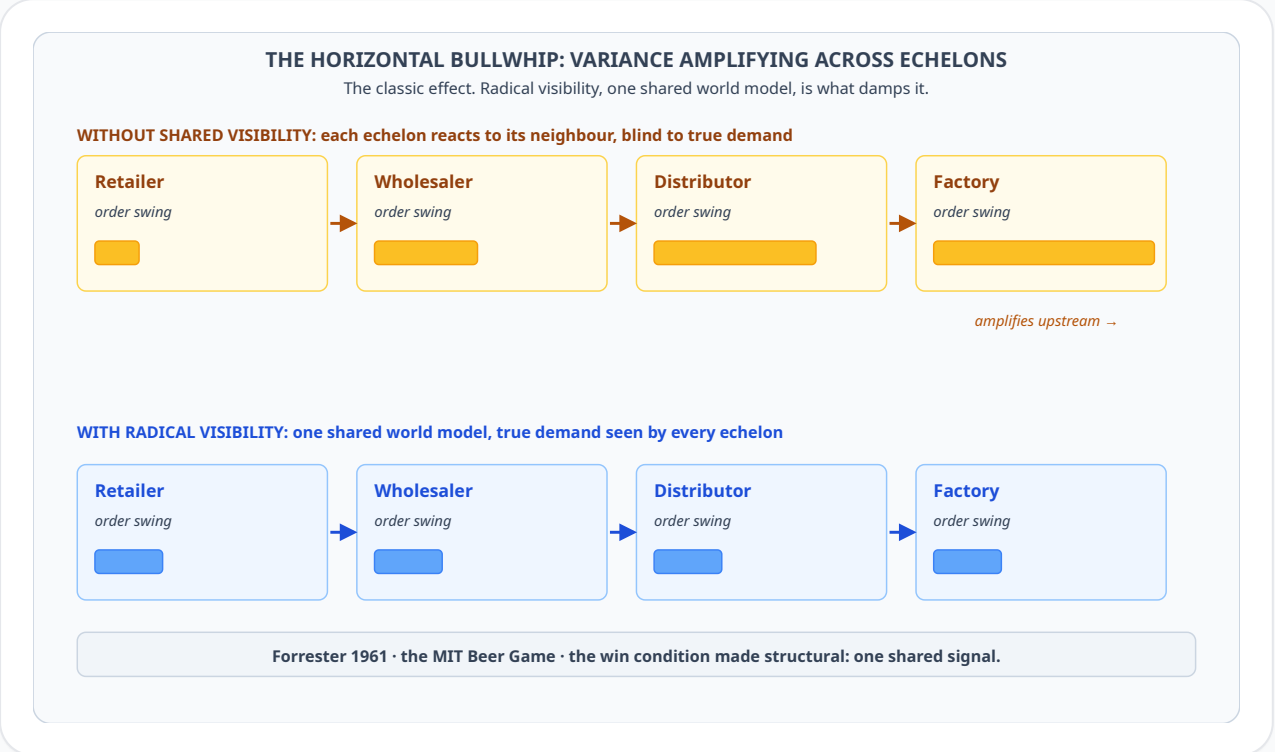
the MIT Beer Game to teach it; Peter Senge made that game famous in *The Fifth Discipline*, turning the effect into a lesson you feel in your hands: four players, each acting rationally, each seeing only their immediate neighbour's orders, produce wild oscillations no one intended. Lee, Padmanabhan and Whang gave the effect its name in 1997. The cause is always the same: every echelon optimises locally because it cannot see true end demand, so it reacts to a distorted proxy, and the distortion amplifies upstream.

That is local optimization and information distortion, Ram's second and third mechanisms, compounding across the horizontal links instead of the vertical ones. And it is exactly what the Beer Game was built to teach: the players lose because they hoard information and optimise locally, and they stop losing the moment they can see and act on the whole.

Autonomy makes that winning condition structural. Every agent reads and writes *one shared canonical state*, a single world model of the network, so each echelon decides against true downstream demand and true upstream commitment rather than a guess. When one agent's decision would touch a parameter another agent owns, the peer-negotiation contract makes it test the change in a parallel scenario first: a counterfactual world in which the network's DAG and the other agents play the change out, producing a candidate plan. Only if that candidate plan improves the *enterprise* scorecard against the current plan does the agent put the proposal to the owning agent. That is what we mean by *radical visibility*: not a dashboard bolted on top of the silos, but a substrate where there is only one version of the signal and every node sees it, and where the objective each node optimises is the enterprise scorecard rather than its own local cost. Senge's lesson, that the players win only when they share information and think in systems, stops being a workshop epiphany

and becomes the default: the Beer Game is unlosable when the information is shared by construction and the goal is shared by design.





Why the higher standard is a good thing

It would be easy to read all this as a complaint: that we are unfair to agents and ought to relax. That is exactly backwards. The higher bar is not a burden we have unfairly placed on agents. It is the right bar, for four reasons.

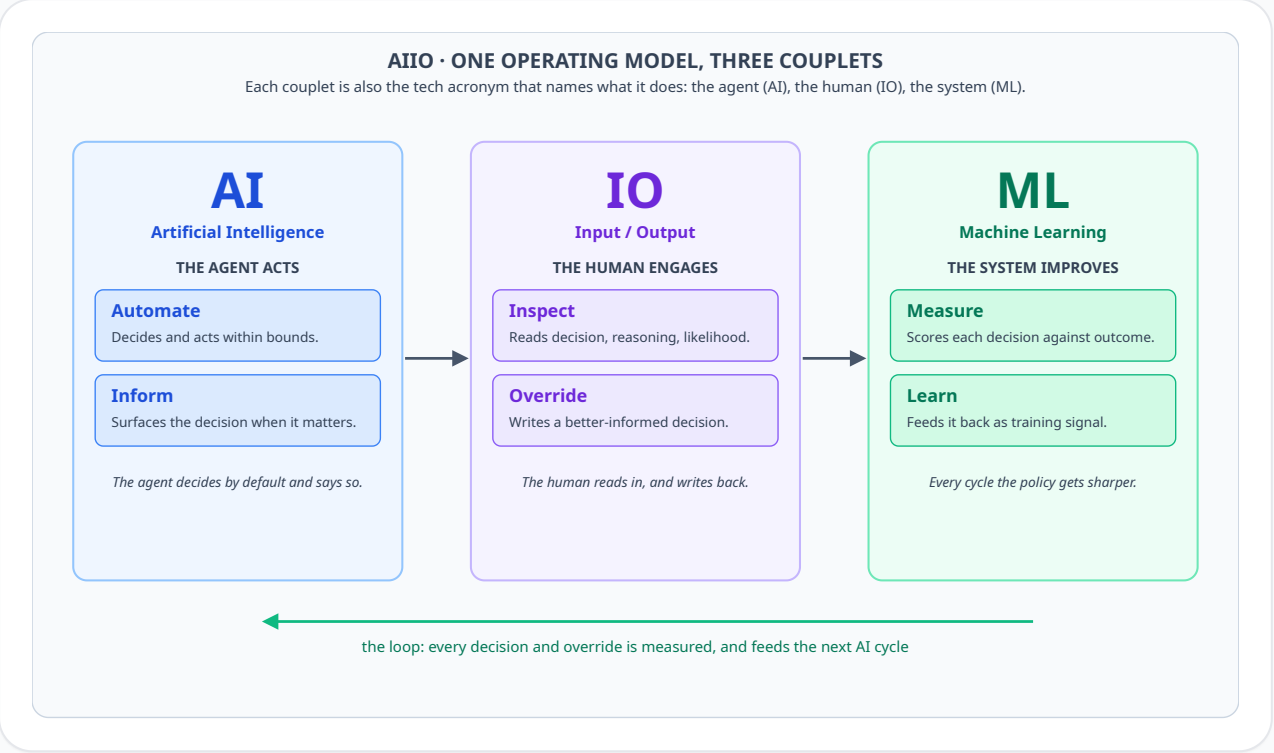
1. It is the only bar a built thing can actually meet. You cannot engineer self-preservation out of a person. You can ask them to rise above their own incentives, every day, unsupported, and the best people often will. But you are asking, not guaranteeing. You *can* engineer it out of an agent. The higher standard is not unfair to the agent. It is the standard that only becomes *available* once a thing is made of code instead of career. We hold a suspension bridge to a higher load than a rope bridge, and no one calls it prejudice against rope.

2. The higher standard is the entire reason to deploy them. If you held an agent only to the standard you hold yourselves, locally optimal, self-protective, the message softened on the way up, it would be no improvement on the organisation you already have. The whole case for the agent is that it clears a bar the organisation structurally cannot. Lower it to the human bar and you have spent the budget to reproduce the Hidden Cost column, only faster.

3. The high bar is the first yardstick the organisation has ever had. This is the one that matters most. The five drifts are tolerated because they are invisible. There is no reference standard to measure them against, which is exactly why the costs stay hidden. Stand an instrumented agent next to the organisation, every decision traced, every estimate carrying a calibrated likelihood, every choice scored against the enterprise, and for the first time

people can *see* what undistorted flow and enterprise-optimal actually look like. Holding agents to the high bar does not leave humans at the low bar. It hands the organisation the measuring stick it never had, and the Hidden Cost column stops being hidden.

4. It is what earns the right to free the human. You can only take a person out of the bottleneck and hand a decision to an agent if you have held that agent to a standard worth trusting. The high bar is precisely what lets a person stop being the approval-relay and become the Override authority, the one role that genuinely needs a human. The standard we demand of the agent is what buys back the human's time for judgement. Hold the agent low and you are back to a human checking every decision, which is the bottleneck Ram's whole article is about.



Agents have their own accidents

None of this means agents are clean. They can miscalibrate. They can act outside the scope you intended. On the narration layer, a language model can say something that is simply not so. If I stopped at “agents do not have the five flaws,” I would be making a boast, not an argument.

The point is the opposite. We hold the agent’s *own* failure modes to the same engineered standard: a conformally calibrated band on every estimate instead of a false point of confidence, a guardrail that holds any decision crossing a bound, a standing human Override, and a hard rule that the language model is never on the decision axis. Same standard, applied to the agent’s accidents and the human ones alike. The difference is not honest versus dishonest. It is *instrumented versus invisible*. An agent’s accident leaves a calibrated, audited trace and trips an alarm. The five human accidents are, by their nature, the ones no one sees until the Hidden Cost has compounded into a crisis.

The standard was always the right one

Ram M closes by naming the work for leaders: design systems that reward transparency over self-protection, enterprise thinking over departmental optimization, truth over comfort, knowledge sharing over dependency, and informed action over excessive caution. Read that again as an engineer rather than a manager. It is a specification. It is, almost line for line, the specification we hold our agents to. He wrote it as the standard human organisations *should* meet and mostly cannot. We took the same standard and made it executable.

So yes, we ask more of our agents than we ask of ourselves. Keep doing it. The higher bar is the only one worth building to, the only one that justifies

the build, and, once an agent is meeting it in plain sight, the first honest yardstick an organisation has ever had to measure its own accidental drift against. The Hidden Cost column was always there. We just never had anything that refused to pay it.

Further reading.

The piece this post responds to.

- Ram M (June 2026). *How Employees Accidentally Cripple Organizations.* LinkedIn: How Employees Accidentally Cripple Organizations. The five-mechanism framing, and the *Hidden Cost* column, this post takes as its starting point, and whose insistence on the word *accidentally* it tries to honour.

Azirella's expression of the standard.

- AIIO: the operating model that lets an agent act by default and keeps the human's override first-class, so neither self-protection nor approval-paralysis is the resting state.
- The Decision Flow Problem: why decision velocity, not headcount, is the constraint, and why local optimization is the tax on it.
- Stop Using Averages: the conformal layer that puts a calibrated band on every estimate instead of a comfortable point.
- Operating Knowledge: the substrate that elicits tacit expertise into a shared, lifecycle-managed store rather than leaving it in one person's head.

- How agents learn: the hash-chained Decision Trace and the override-to-training loop that keep the record honest and the system improving.

See Autonomy in action

Walk through how Autonomy models, executes, monitors, and governs supply chain decisions with autonomous AI agents.

See It Live