

The Hybrid Enterprise

An operating layer for the firm after the Coasean inversion

*For the firm that means to stay coherent at machine pace, keep its memory callable,
and change its own boundary on purpose.*

A white paper for enterprise architects, software engineers, and executive leadership.

Executive summary

Enterprise AI is not failing on capability. It is failing on coordination. This paper is written as a detective story, because market forces are making a killing.

There is a body on the floor. Two point six trillion dollars walks into enterprise AI this year, and in most firms nothing measurable walks back out. Evidence already gathered from independent witnesses who do not know each other, and whose numbers converge anyway. Their accounts of what's happened, and what continues to happen, line up. They are credible.

There is a vital clue among their accounts. The *same models* and the *same vendors* are producing radically unequal returns. That's a signal. The differentiating variable is not the technology.

There are suspects with alibis, nine of them, each a familiar response that is true and good work, but partial, and aimed at the wrong constraint. And there is a killer, identified by name: an inversion. AI collapsed the cost of doing. It did not collapse the cost of deciding, aligning, and answering for outcomes.

The dominant design of work changed. It took on a new shape, a new form. The constraint moved, and the enterprise is still organized with the same one that's held for decades. And the fixes that raise local performance are the ones every advisor sells and every rival installs, both in good faith, so they buy motion, not position. Position is not on the shelf.

I crossed paths with this killer before. And again recently at close range. So part of this paper is testimony, however we will stand on objective evidence along with a verifiable claim before we are done.

Years before the current queue crisis in open-source security, I ran exactly that same loop by hand for a federal healthcare program. It held because a person held it. That is the pattern in miniature: the coordination was never a system.

The repair Monday is not a bigger model, another pilot, or a new committee. It is a missing layer of enterprise structural architecture. Enterprises run a *runtime* for compute. Nothing runs one for coordination, yet. This paper specifies that structure, *the coordination runtime*: six primitives in three pairs (Intent and Policy, Roles and Records, Cadence and Signal), composed across five functional "floors," a familiar metaphor, not an org chart, where people partner with AI for machine capacity, carried on a Stewardship Spine that shares intent down and carries metrology for telemetry up.

The claim is testable, and this paper stakes it: firms that build the operating layer while deploying AI will reach the productivity harvest years before firms that deploy AI alone. Two public baselines from 2026 are named for re-measurement in 2028. If the gap fails to appear, this framework is wrong, and its author will say so in public. The paper closes with a diagnostic an executive team can run in one meeting and join the cohort. There is an order of operations in the first ninety days of the build, then it carries and compounds. You get a virtuous cycle.

1. The body on the floor

Begin where a detective begins: something is wrong, and it is not subtle.

Investment is not the constraint.

- Gartner forecasts worldwide AI spending of 2.6 trillion dollars in 2026, up 47 percent in a single year.¹

Adoption is not the constraint.

- McKinsey's late-2025 global survey finds 88 percent of organizations using AI regularly in at least one function.²

Return is the constraint. The same McKinsey survey finds just 39 percent reporting any EBIT impact at the enterprise level.

- MIT's Project NANDA, examining custom enterprise GenAI pilots through mid-2025, found 95 percent showing no measurable P&L return, and located the failure not in model quality but in a learning gap: systems and organizations that do not retain feedback or adapt.³ - S&P Global found the share of companies abandoning most of their AI initiatives jumped to 42 percent in 2025, from 17 percent a year earlier.⁴ - Gartner predicts more than 40 percent of agentic AI projects will be canceled by the end of 2027.⁵ - The Stanford AI Index 2026 reports that 89 percent of enterprise AI agent deployments never reach production.⁶

Different investigators. Different methodologies. Different denominators. The numbers converge anyway. The technology works in the demo and stalls in the enterprise, and the stall has the same shape everywhere.

The value gap · money in, use up, return missing

Investment · not the constraint

\$2.6T

worldwide AI spending
forecast for 2026

up 47% in one year

Gartner, May 2026

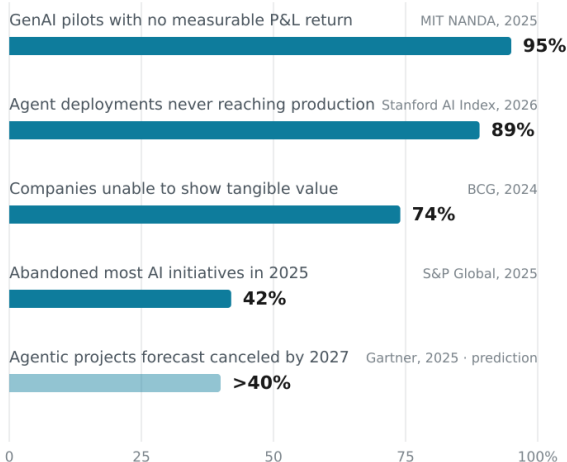
Adoption · not the constraint

88%

of organizations use AI regularly
in at least one function

McKinsey Global Survey, Nov 2025

Return · the constraint



Five independent measurements of the same missing return. Percentages describe different populations (pilots, deployments, companies, projects); the bars share an axis, not a denominator. The lighter bar is a forward-looking prediction. Sources at notes 1 to 9 of this paper. Schematic, not to scale.

Figure 1. The value gap, carried in three: money in, use up, return missing.

Here is the anomaly, and it is **the clue** this whole case turns on. Davenport and Bean find 39 percent of Fortune 1000 firms with AI at production scale, up from under 5 percent two years prior, so the ceiling is real and some firms are through it.⁷ PwC's 2026 study of 1,217 executives found 74 percent of AI's economic value flowing to just 20 percent of organizations, and named the discriminator: workflow redesign over tool layering, with technology delivering roughly 20 percent of an initiative's value and redesigned work the other 80.⁸

Same models. Same vendors. Radically unequal returns.

"Same models" means the same underlying LLMs: the frontier models themselves, the GPT series, Claude, Gemini, Grok, Llama. Any Fortune 1000 firm and its worst performing competitor are drawing from the identical short menu of foundation models, often literally the same model version behind the same API.

"Same vendors" is one ring wider. The inner ring is the model providers themselves: OpenAI, Anthropic, Google, Meta, xAI. The outer ring is the enterprise software vendors who resell those same models inside their products: Microsoft Copilot, Salesforce, ServiceNow, and the rest of the SaaS shelf, mostly the same handful of foundation models wearing different badges. The outer ring matters because an executive who says "we don't buy from OpenAI, we buy from Microsoft" is still inside the argument.

If the technology were the cause, the returns would track the technology. They do not. They track something else, something organizational, and a reader of Deming already knows where to look: he estimated in 1986 that

94 percent of variation in performance belongs to the system, the responsibility of management, and 6 percent to special causes.⁹ Nor is the ratio a relic.

The witnesses above just re-measured it without naming him: PwC clocked technology at 20 percent of the value and redesigned work at 80; NANDA located the failure in the learning loop, not the model. Same verdict, new denominators. And a system is not a given. It is designed, or it is inherited unexamined. The enterprise environment itself is a system and it's our crime scene. The rest of this paper walks it.

2. Testimony: I have stood inside this before

Every case benefits from an eye witness who was in the building.

In May 2026 the open-source security world hit a wall that made the news: an AI research system found loads of critical vulnerabilities faster than the people owning them could verify each and ship fixes for them all. Anthropic's own update reported that of an estimated 530 high- and **critical-severity flaws disclosed** to open-source maintainers, only 75 (14.2%) were patched at the time of the update, and maintainers asked for slower disclosure.¹⁰ Read that again for a moment more. The machines found the critical flaws, and the people responsible *asked them to slow down*.

That is not a good practice. That is also not resistance. *That is a system reporting its own cadence*. And I recognized it immediately, because years back I ran that loop myself.

I led enterprise security patching for a federal healthcare program. The work had a shape, one I had to conceive before I could run it. Intent came down: reduce the risk posture, these classes of vulnerability, this window. I did not set that; I received it. Then I triaged every critical and urgent patch against that intent. Then I sequenced implementation across every environment, inside a month, around adjacent initiatives, do-not-disturb windows, and people who could not be interrupted. Added a catch-all event, absorbed what the standard cadence could not hold. After every batch, a report went back up: progress, posture, exceptions, the new flaws and complications that surfaced mid-flight with new context. Then it started again.

What ran underneath all of it. What gets fixed first? Who owns the call? What rhythm it ships against, and how the signal comes back? Translated as three pairs that's:

- Policy and Intent. - Roles and Records. - Cadence and Signal.

Not three steps in a line, a loop. It never closed. It ran continuously. Fit for purpose.

At a trickle, one capable operator can hold a loop like that together, given enough time. I did. And that is exactly why the person (and the team), not the structure, looked like competence and discipline. It was enough for its time. People good at the job, earning trust by showing up turn after turn. It was a person standing in for a system that did not exist. I was the buffer, and at that volume the buffer held.

Now multiply the find rate by just ten (one order of magnitude) and the stand-in fails. Not because the operator got worse. Because the loop was never built to run without a hand on it. The hand buffered for the missing structure in the enterprise architecture.

One more piece of testimony, from the other side of the ledger and a different industry: utilities and energy.

The best solution architect I ever worked with once conceived and shipped a working prototype in an afternoon. It took another three weeks for the client and the team to agree on what they had built and what would make it stick. The doing took hours. The aligning took twenty-one days.

Today every enterprise and many ICs can produce the afternoon prototype. Almost none owns anything that shortens the three weeks.

So hold both stories. A queue that scaled ten times while its clearing cycle did not. A prototype that arrived in hours and waited weeks for agreement.

Those scenes are not drawn from different eras; they happened within twelve months of each other, one operator, two seats, in two industries as unlike as federal healthcare and utilities and energy. The proximity is the point: same year, different sector, different function, and the shape appeared anyway.

Two sightings that close together could still be one person's unlucky assignments. But the scene that opened this section is the third sighting, years later, at global scale, in systems I never touched. Once is happenstance. Twice is coincidence. The third time, as Ian Fleming's Goldfinger put it, is enemy action.

For us, that means whatever killed the returns in section one was operating long before the current models shipped, and it does not care which industry.

The inversion is one of two pressures, and they are easy to confuse. The first is vertical. The binding constraint moved up, from production to coordination, and most firms still work the floor below it. The second is lateral. In 1973 the biologist Leigh Van Valen named the Red Queen¹¹: in a system where every player adapts at once, a species can improve continuously and hold the same relative position, because its rivals improve too. Enterprise AI is built for this. A capability that diffuses, the shared copilot, the same retrieval stack, raises local performance and returns nothing relative, because every competitor installs the same thing and the market resets its expectation around it. The vertical error works the wrong constraint. The lateral error works a common one.

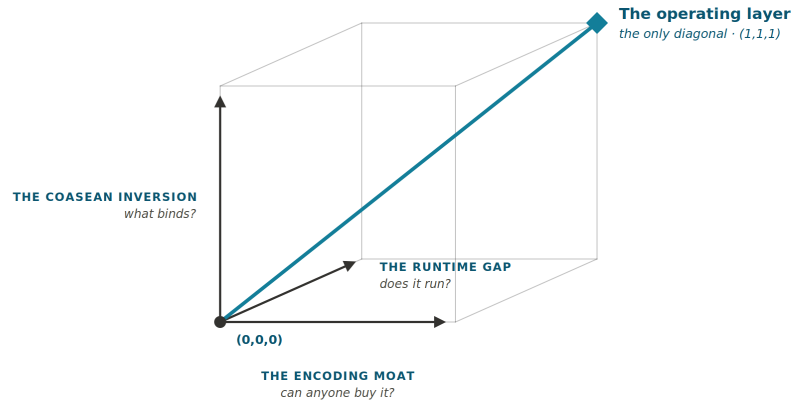
A third question sorts whatever survives the first two: does it run? Continuously, at the pace of the work, or only at design time.

The operating layer answers all three, and it is the only response that does. Coordination encoded to a firm's own intent is the one capability that does not ship in a competitor's release notes. The carrier is rentable. The encoding is not. That is what turns motion into position.

THE POSITIONING CUBE

Which constraint · Can anyone buy it · Does it run

Three questions on three edges. The **operating layer** is the only move that **answers** all three.



READING THE EDGES

Each edge runs from the old condition at the origin, the org chart (0,0,0), to the operating layer (1,1,1).

THE COASEAN INVERSION

what binds?

Doing
draft, build, execute

→ **Deciding**
plan, coordinate, review

THE ENCODING MOAT

can anyone buy it?

Common
anyone can rent it

→ **Encoded**
yours, uniquely encoded

THE RUNTIME GAP

does it run?

Design-time
snapshot, point in time

→ **Runtime**
standing, continuous

Figure 2. The positioning cube: which constraint, can anyone buy it, does it run. The operating layer is the only diagonal that answers all three.

3. The usual suspects

Nine responses now stand in for the repair, and every one of them has an alibi. Hiring a Chief AI Officer as the answer. Change-management programs. Maturity ladders. Use-case portfolios. Tooling portfolios. Governance modules. Org redesigns. Leadership rewiring. Pilot-and-scale. Interrogate each and it checks out clean: none is wrong, each is true work, each has genuinely improved some firm somewhere. That is what makes them suspects worth clearing carefully. We keep arresting them for this crime, and the crime keeps happening.

BCG's survey of one thousand executives already read them their rights: 74 percent of companies were unable to show tangible value from AI, and the difficulty was attributed roughly 10 percent to algorithms, 20 percent to technology, and 70 percent to people and process.¹² Somewhere, Deming nods.

Work the lineup in that order. The algorithms were not the problem. The technology was not the problem. The coordination of people and process was, and none of the nine responses runs coordination continuously.

That is the tell that clears them. A role without instruments is a mandate. A pilot portfolio without records is a set of episodes the organization will not remember. Each of the nine optimizes something locally, but it does not move the constraint. And it does not compound value until we re-organize the enterprise system to do so. Goldratt priced this arithmetic in 1984: an hour saved at a non-bottleneck is a mirage.

The test for any proposed response: does it run continuously, does it carry intent down and metrology for telemetry up, does it leave a callable record? If not, it is a better version of the old answer, and more of the old answer is not what this case needs.

So the suspects walk. What remains is to name the root cause. What moved the constraint we all feel? Whose shadows have we been glimpsing?

4. The killer: what inverted

Ronald Coase asked in 1937 why firms exist at all, and answered: because coordinating inside the firm was cheaper than transacting outside it. The boundary of the firm sits where the two costs cross. Everything about the twentieth-century enterprise, the hierarchy, the headcount, the org chart, is downstream of that crossing point. A strategically competitive decision point, but also fundamentally an economic decision point to run operations profitably in the black instead of the red.

In 2025 a group of economists asked what happens when AI agents drive transaction costs toward zero, and titled the question "The Coasean Singularity?"¹³ The question mark is theirs and it belongs there: whether firms unbundle into markets is a long question with a long horizon. Howard Yu at IMD carried the term to practitioners, arguing that AI's collapse of coordination costs undermines the economic rationale for the org chart itself.¹⁴

An operating executive does not need to wait for the long question. A nearer fact is already on the floor, and this paper names it the Coasean inversion.

AI collapsed the cost of production: the drafting, the coding, the analysis, the execution of bounded work. It did not collapse the cost of coordination: deciding what to do, aligning the people and systems that do it, and answering for the outcome.

A fair skeptic will note that search, negotiation, monitoring, and enforcement are the costs that move firm boundaries, and that those fall only as agents become trusted to act. The skeptic is right, and the distinction is the point. The *doing* got cheap first. The *deciding* did not. The constraint moved.

The Coasean inversion • the binding constraint moved from production to coordination

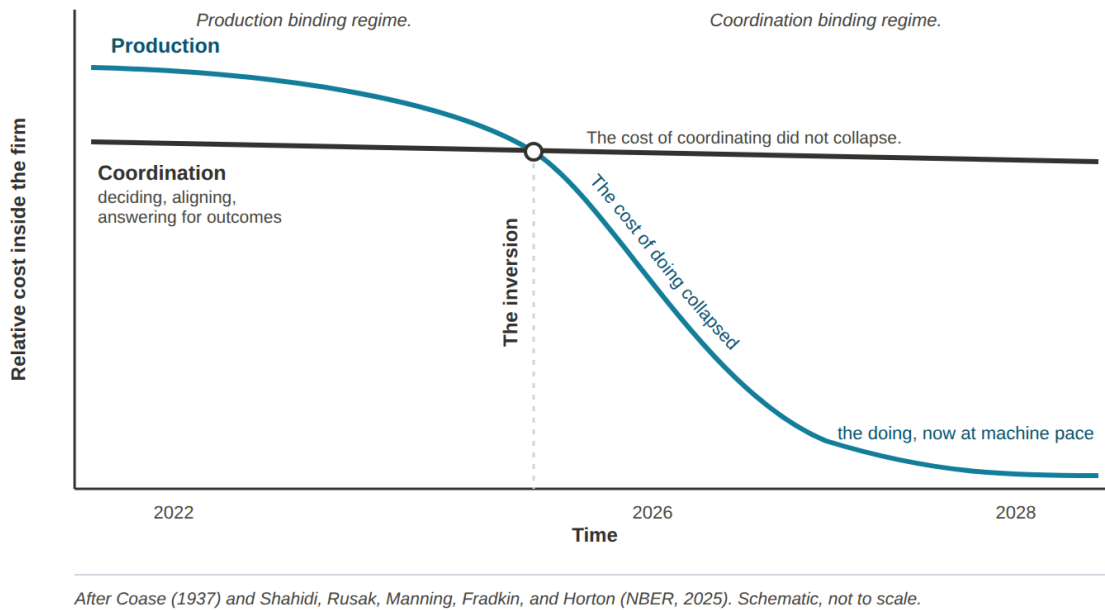


Figure 3. The Coasean inversion.

Now the motive and the method, because a name is not an explanation. Coordination never appeared on an org chart because it never had to. It was absorbed by people. Humans carried the tribal knowledge, translated between systems that defined "customer" three different ways, negotiated the unwritten rules in conversation. Humans were the buffer, and at human pace the buffer held. The org chart described positions and the operating model described the design, and the space between them was staffed invisibly by whoever was willing to compensate. My patching program was one instance; your firm runs on hundreds of them.

AI removed the buffer's cover in both directions at once. Machine-pace execution generates decisions, exceptions, and drift faster than any hallway or approval queue can absorb. And the deterministic assumption under legacy governance broke: generative systems do not behave the same way twice, so trust that was once installed at deployment now has to be maintained continuously, the way uptime is.

This exposes a distinction most enterprise architecture has never had to make.

The org chart is a design-time artifact.

The operating model is a design-time artifact.

Both are snapshots, silent between revisions. What the enterprise now lacks is a runtime: the standing structure that carries intent to the work, reads what is actually happening, and corrects while the outcome is still cheap to change. Enterprises already run compute, networks, and security as continuous disciplines with telemetry, on-

call rotations, and error budgets. Coordination, the thing that determines whether all the rest compounds, still runs on quarterly meetings and personal heroics.

Name what is missing precisely: a coordination runtime. Infrastructure already talks this way. A container runtime is the thing that runs containers, and nobody asks for the preposition twice. The unfamiliar part is the promotion.

- In *at runtime*, runtime is a *when*. - In *the coordination runtime*, it is a *what*: a standing thing the firm operates, not a moment the work passes through.

The promotion from when to what is the repair this paper specifies.

The operating layer is the coordination runtime.

Don't stop at the next architectural review. That's another design time practice. Encode your firm's coordination runtime then examine it on cadence and service it. Maintain it. Optimize it for your niche, function, portfolio and the whole business.

One more consequence, and it answers the obvious objection. If markets are getting cheaper, why build anything inside? Because the inversion binds on both sides of the firm boundary. Make-or-buy is no longer an annual decision; capable alternatives now arrive monthly, and much of what enterprises paid SaaS premiums for in 2025 ships natively from model providers in 2026.

The market will even rent you the deployment: forward-deployed engineers who embed, encode your workflows, and leave with the badge. So the question is no longer build versus buy the model. It is whether the encoding stays when the vendor does not.

A firm without a coordination runtime cannot even see those crossings, let alone act on them safely. The operating layer is not a defense against the market. It is the instrument that lets the firm consume the market deliberately, rather than be consumed by it: fund, keep, combine, deprecate, or subscribe, with the reasoning recorded and reversible.

The firms that thrive after the inversion will not be the ones that build the most. They will be the ones that can change their own boundary on purpose.

The case is made. The killer is named. From here the story changes genre, because knowing who did it is worth nothing to the executive still standing in the building where it happened. What follows is the repair, and the repair is yours to run. My role from this point is the one I earned in the patching years: the guide who has walked the building before, handing over the map.

5. The building: a reference architecture

The Hybrid Enterprise is an archetype, not a maturity stage. It is the settled form of the firm in which people and machine capacity work paired, and it sorts the enterprise by decision horizon rather than rank.

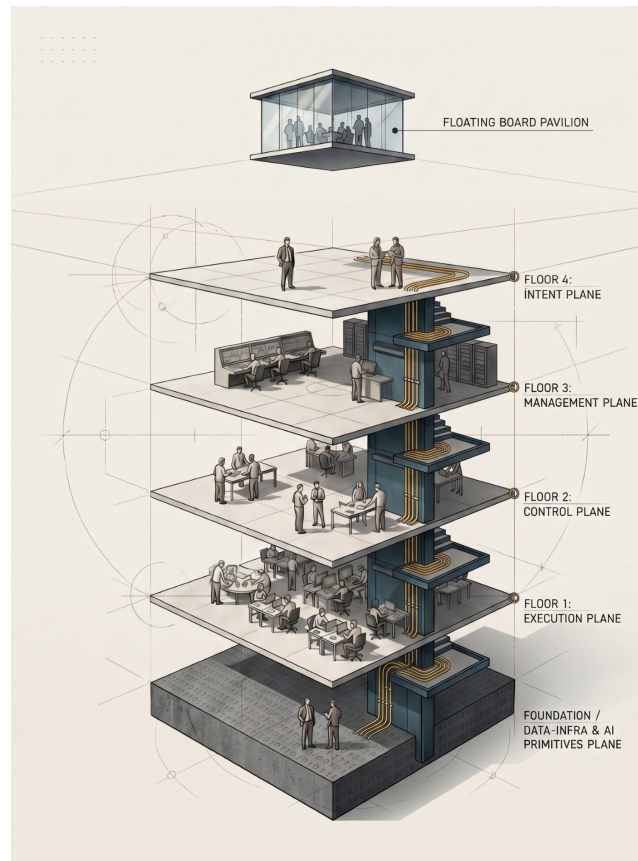


Figure 4a. The Hybrid Enterprise as a cutaway: the floors as planes of work. Labeled planes in this rendering; the paper calls them floors.

Five floors carry the work, each a kind of work, not a job title. And a floor is not headcount. One floor can be held by a team, by one person, or by one person paired with machine capacity doing what used to take several hires. The pairing moves headcount, not accountability. Five floors are five kinds of decision, not five layers of management. The lone entrepreneur makes all five.

F0, Foundation: data, infrastructure, networking, and the AI primitives themselves. The ground the floors rest on. The people here add value by staying off the live loop: building, tuning, and refitting the primitives everything above runs on. Foundation elements emerge, get built, combine, collapse, and get deprecated on their own lifecycle.

F1, Execution: the floor of the nearest horizon. Machine capacity clears the inbound work at pace; the practitioner receives the exception with the triggering rule, the recommendation, and the cost of being wrong

already on the table. The human owns the call.

F2, Control: the floor of rhythm. Managers tune thresholds, grant and revoke runtime authority, and catch drift while it is still in-policy. The hard case is precisely the one where nothing alarms: an agent acting within its rules while the ground shifts under them.

F3, Management: the floor of the portfolio. Leaders read across functions, where each function's dashboard is green and the composite is not. They prune: folding down capacities the market has outrun before they harden into unquestioned cost. And they refuse the single yardstick. A core business, an emerging bet, an option: each answers a different question, and only the first answers to ROI.¹⁵

F4, Intent: the longest horizon. The C-suite codifies what the firm is here to do, places the few and final bets, and owns the decision to deprecate. These decisions are the farthest reaching and the least reversible. Machine capacity reads the world and war-games the options. It cannot place the bet, because accountability does not transfer. You can build or buy capacity. You cannot buy the seat.

Above the floors, set apart and never numbered, sits the Board Pavilion: fiduciary oversight, watching whether the work can be trusted. The floors answer for the work; the pavilion answers to the absent, the people whose capital and risk ride on work they will never watch run. That is why it has no number.

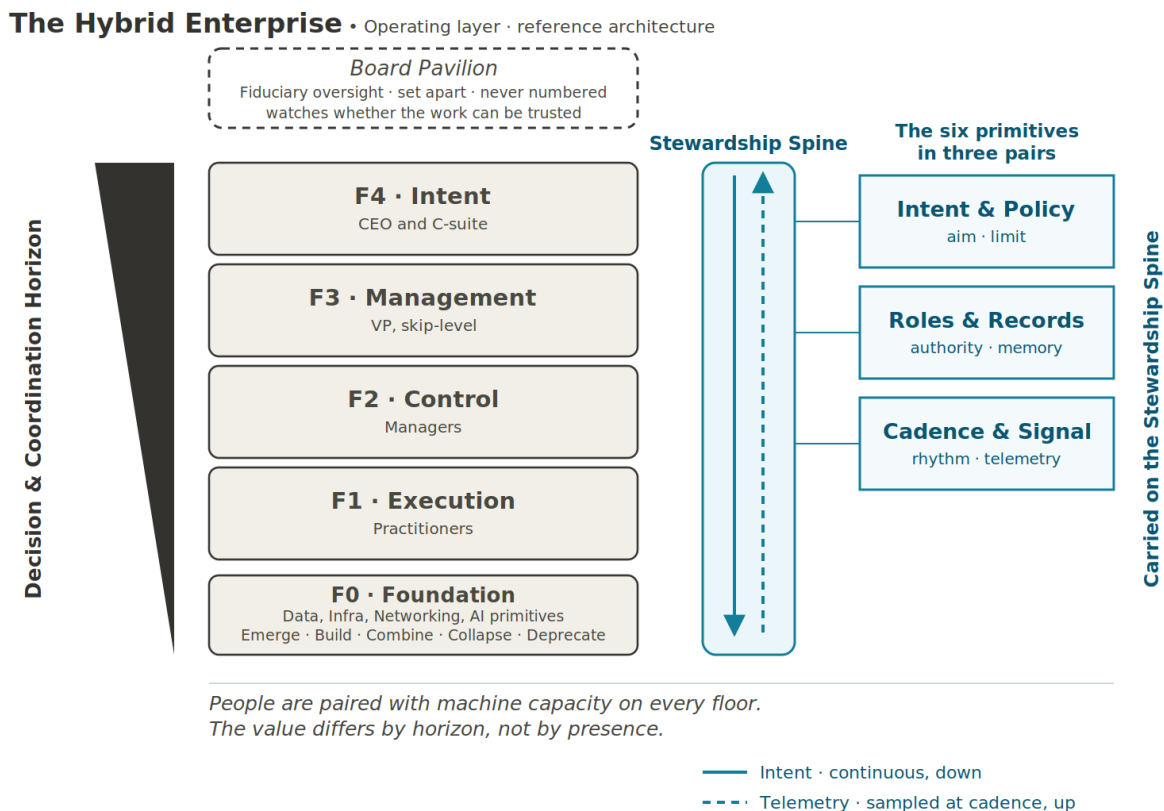


Figure 4b. The same building as functional abstraction: the reference architecture.

Two properties make the architecture load-bearing rather than decorative.

People are paired with machine capacity on every floor. The value of the human differs by horizon, not by presence. This is supervised autonomy: the human on the loop, not in it. No floor signs off on the floor below; the structure carries intent and signal, not stacked approvals. Decades of human-factors research say the alternative fails predictably: humans reviewing every machine decision go numb, and the signature stops carrying judgment long before anyone notices.¹⁶

The pattern is already visible at the level of CEO statements. Google reports more than a quarter of its new code generated by AI, "then reviewed and accepted by engineers." Microsoft estimates 20 to 30 percent in its repositories.¹⁷

The production moved to the machine. The acceptance stayed human. That division of labor is the Hybrid Enterprise in miniature, and Klarna's much-cited course correction makes the same point from the failure side: after replacing the work of seven hundred agents with AI on cost grounds, its CEO conceded that quality had suffered and rebuilt around AI for routine volume with humans on the work that carries the brand.¹⁸

The Stewardship Spine runs the full height. Intent travels down, so the person meeting reality knows what the reality is for. Telemetry travels up, sampled at cadence, so the floors above act on what is happening rather than what was true last quarter, month, or week. Stewardship runs alongside both, and it informs rather than enforces. That separation of concerns is deliberate; it avoids a conflict of interest.

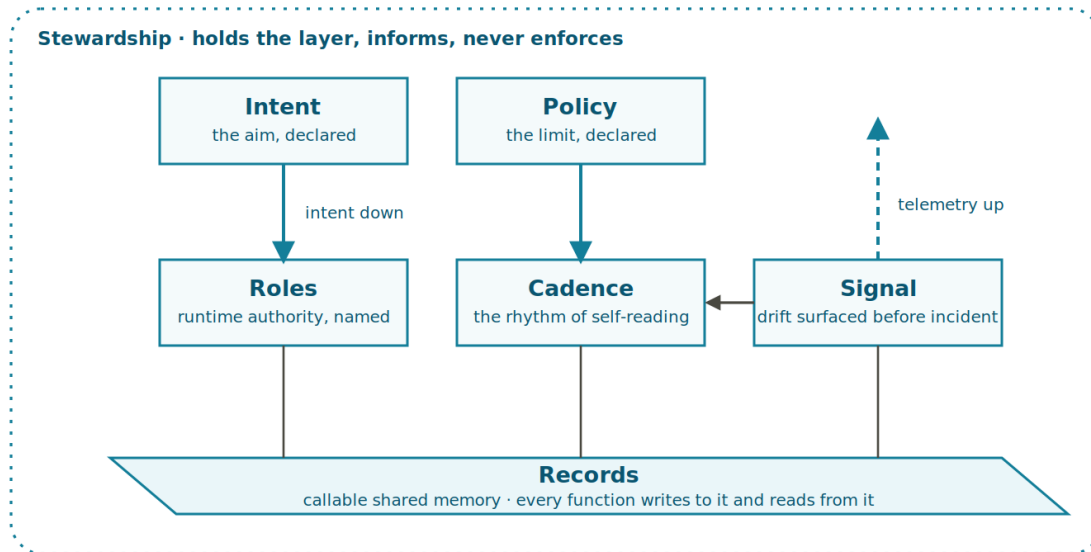
Notice what the architecture is doing to my patching story. Intent came down to me; I carried it by memory. Telemetry went up from me; I assembled it by hand each cycle. Authority, sequence, and rhythm lived in one operator's head. The floors and the Spine are that same loop with the person taken off the crank and put in charge of the design. Nothing about the work changed. What changed is that it no longer depends on the hand.

6. The operating layer: six primitives

The runtime the enterprise is missing decomposes into six primitives, in three pairs. Software will carry them, and must: the closure test at the end of this section requires each primitive to be operable at machine pace. Records in particular is only real as software, memory made lossless and callable instead of carried by hand, lossy at each retelling.

What cannot be purchased is the layer itself. Tooling arrives from the market; the discipline, the firm's own recorded decisions, and the patterns learned from its own operations do not. Those exist only after the loop runs, and they are the part that compounds. Buy or subscribe where the make-or-buy math says so; the quarterly question exists for exactly that. What no vendor can sell is your memory.

The operating layer • six primitives composing at runtime



Three pairs. Failure signatures: sprawl and incoherence (Intent & Policy); audit and amnesia (Roles & Records); drift and incident (Cadence & Signal). Diagram form after Beer's Viable System Model lineage.

Figure 5. The operating layer.

Intent and Policy. The codified what-we-are-here-to-do and the codified what-we-will not-allow. The aim and the limit. When these are absent at the layer where the firm should be coherent, every team improvises its own. The failure signatures are sprawl (no policy) and incoherence (no intent).

Roles and Records. Roles: who holds runtime authority, by name, at each floor. Records: the callable shared memory of decisions with their basis, the conditions that held, the alternatives rejected. A record written for audit goes to a drawer. A record written for downstream use stays alive: the next person, the next cycle, the next

agent can invoke the reasoning rather than reconstruct it. The failure signatures are the audit trap and organizational amnesia.

Cadence and Signal. Cadence: the rhythm at which the system reads itself, chosen deliberately, coupled to machine pace where stakes demand it and deliberately uncoupled where they do not. Signal: what surfaces drift before it becomes an incident. A review cycle without a signal protocol is a calendar event. The failure signatures are drift and incident.

The pairs compose. Signal feeds Cadence. Cadence pressure-tests Policy. Policy clarifies Intent. Intent reshapes Roles. Every function writes to and reads from Records. That composition is why partial answers do not add up to the layer: a role without records is a hero, records without cadence are an archive, cadence without signal is ritual.

One objection arrives here reliably, usually from the top of the house: we already have all six. True, and so did every firm in section one's failure statistics. You have them as artifacts, the strategy, the org chart, the minutes, the quarterly review, built for human pace and silent between revisions. The layer is not six new possessions; it is the same six, running. One question settles which one you own: if an agent acting within policy drifted off intent last Tuesday, what in the firm would notice before the quarter closes, other than a person getting lucky?

A word of intellectual honesty the field rarely offers. This set of six emerged from practice and converges with older systems thinking, most directly Stafford Beer's Viable System Model and the cybernetic tradition of Ashby; the structural diagnosis has been independently reached by others, notably Stefan Norrwall's Axion model of organizational coherence.¹⁹ The six are a working set with a closure test (a primitive must be operable at machine pace and necessary for the legibility of intent and judgment at runtime), not a proven complete decomposition. Enterprise architects should treat it the way they treat any reference model: as a starting frame that earns its keep in use.

7. Stewardship: the posture and the mechanism

Governance answers a different question than the one now being asked. An audit tells you where you were. The operating layer tells you where you are. Both are necessary. Only one of them is new work. Stewardship is the discipline that shapes and sustains the operating layer. The word has a lineage, borrowed on purpose.

Clinical medicine named algorithmic stewardship in 2020, when a drifting model was already a patient-safety problem.²⁰ The same year, ML theory formalized the loop underneath: deployed models change the behavior they predict, and the change forces retraining.²¹ Medicine built the committee. The theorists built the mathematics. Neither built the enterprise discipline. This paper generalizes both. Three commitments define it.

It is a posture, not a role, and it is taken, not assigned. It does not require a Chief AI Stewardship Officer, and it does not belong to any single function, not technology, not HR, not risk. Centralizing it defeats it: seated with power it becomes enforcement, seated without power it becomes theater.

It distributes across the floors, adopted by people who are already there. At F1 it looks like authorship of the exception. At F2, instrumentation. At F3, portfolio judgment. At F4, fidelity: insisting the Spine carries real signal, not curated signal. The patching loop ran on one hand. This runs on many, by design.

It informs rather than enforces. Enforcement polices the turn; stewardship shapes the machine that turns. The distinction is what keeps the layer from becoming the bureaucracy its critics expect.

It runs on two cadences that already fit the corporate calendar. Monthly, at the operational level: the owners of each AI-bearing capability report against a consistent baseline frame.

- Is this still aligned with intent?
- Has context shifted?
- Has how people use the capability shifted?
- What approaches deployment that needs attention before it ships rather than after?

The third question is the one no existing literature asks. ML theory treats behavior change as noise to cancel. Interface research treats it as risk to design away. Stewardship reads it as demand: the pair generates the signal, and changed behavior is the earliest telemetry of a need the roadmap has not caught.

Quarterly, at the portfolio level, as a standing segment of an existing review: one person with cross-enterprise visibility and access synthesizes for the leaders who set strategy, enable it, and fund it. The standing questions are strategic: What do we fund, keep, uplift, combine, deprecate, or subscribe to? Has the make-or-buy math shifted this quarter?

And one transition no calendar catches unless it is named: release, where spend changes character, from time-boxed project money to operating support plus variable consumption. That handoff is pre-defined, owned by a

named role, and written to Records before the capability ships, or it is inherited by whoever notices first. Legacy support models were built for deterministic software; they inherit none of this.

Two gauges keep the whole system honest, read together, per floor. The governed ratio: how many agents run per human at this altitude. The override rate: whether the human judgment is still real.

A rising ratio with a healthy override rate is maturity, and the higher ratio is earned. A rising ratio with an override rate collapsing toward zero is automation bias wearing a green dashboard.

The safe ratio is relative, learned, and dynamic; make it a target and it stops measuring anything.

And read the gauges whenever autonomy changes hands.

Graduating an agent from in-the-loop to on-the-loop is not a dial a team quietly turns because the agent seems to be performing. It is a decision, written to Records, read at Cadence, reversible on evidence.

Return once more to the scene from section two, because it is the gauges' proof in the wild. Anthropic's frontier-model security research found flaws in open-source software far faster than maintainers could absorb: 530 disclosed, 75 patched, and a request to slow down.¹⁰ Detection scaled past remediation, which is still a human job. Finding stopped being the limit; deciding, sequencing, and owning became it. That is the inversion in one incident report, and it is not a security story. It is the general condition: wherever machine capacity raises the find rate, the constraint moves to the coordination the old cadence was never built to carry.

The queue does not ask whether you have built the layer. It just arrives, then waits on you and me until we build and encode the layer. Then it serves us where, when and how we choose to design it to do so.

The regulatory runway points the same direction. The EU AI Act's high-risk obligations, provisionally deferred to December 2027, and the automated decisionmaking rules taking effect in several U.S. states, California's among the earliest, converge on the same requirement: consequential machine decisions must be explainable, attributable, and auditable.²² Built as an operating-layer property during the runway, that capability is cheap and compounding. Scrambled as a compliance project at the deadline, it is expensive and brittle. A runway is for takeoff, not parking.

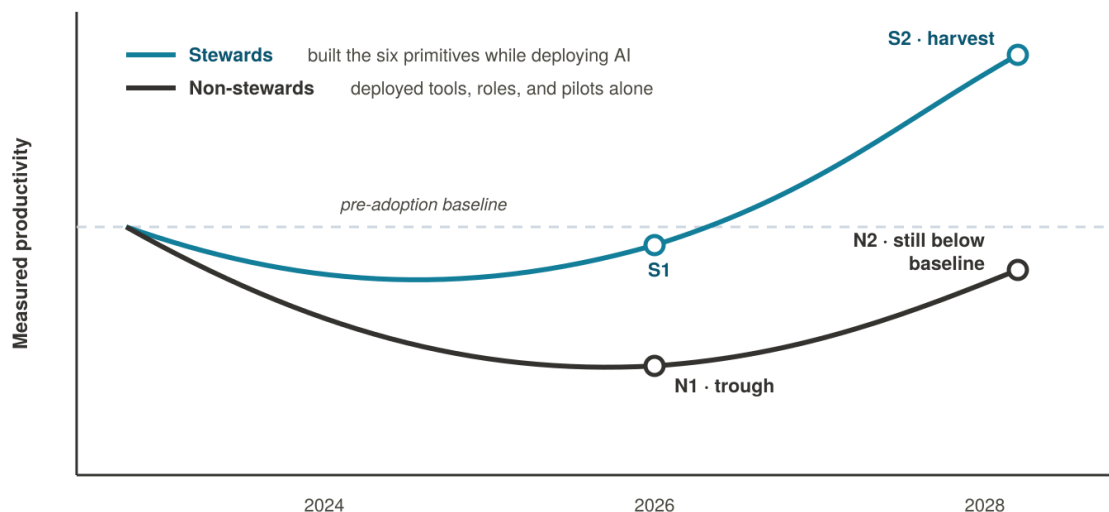
8. The skeptic's page: a falsifiable claim

A good detective shows the court how the case could fall apart. Frameworks in this territory rarely accept the risk of being wrong. This one does.

The productivity J-curve describes why general-purpose technologies depress measured productivity before raising it: the complementary intangibles, process redesign, retraining, and restructuring, are expensed before the payoff compounds.²³ The operating layer is precisely that complementary intangible for enterprise AI.

The claim: firms that build the six primitives while deploying AI (stewards) will clear the trough and reach harvest measurably earlier than firms that deploy AI while working only the nine familiar responses (non-stewards).

A falsifiable claim—two cohorts on the productivity J-curve, re-measured in 2028



Curve after Brynjolfsson, Rock, and Syverson (2021). 2026 anchors: Stanford HAI 2026 AI Index; Davenport and Bean, MIT Sloan Management Review (2026). Both readings re-measured in 2028. Schematic, not to scale.

Figure 6. Two cohorts on the J-curve.

Two public 2026 baselines anchor the test: the Stanford AI Index production-gap reading and the Davenport and Bean production-scale rate.²⁴ Both will be re measured by their publishers around 2028.

And here is the alternative verdict, stated plainly so no reviewer has to state it for me. It is possible that the spread in section one is explained by something this paper underweights: sector exposure, data readiness, balance-sheet room to endure the trough. It is possible the six primitives are the wrong decomposition, or an incomplete one; that is flagged as open in section ten. What is not possible, on the evidence assembled here, is that the gap is technical. The models are the same models. If the gap between coordination-invested firms and

tool-invested firms fails to appear in the 2028 readings, this framework is wrong, and its author will say so in public.

9. Monday, and the first ninety days

The reader this paper is for does not need a verdict. You need a first move, and it fits in one meeting. The scope is yours: the whole portfolio, one function, or the single loop that already worries you. The question is the same at every size.

Ask one thing in that meeting: can someone in this room account for every portfolio component where AI is in play, and for how it holds together? The answer measures the layer. An account the room owns, one you can then stress with the two gauges, the governed ratio and the override rate, means the layer exists. No account, or one that lives in a single head, means it does not. When I held the patching loop, the honest answer in that room was a name, mine, and no one had noticed that the name was the system. Ask whose name it is in your room.

Then the build, smallest real version first.

A word on why this order, because the sequence is dependency, not doctrine. Roles come first because they are the only primitive that costs nothing but a decision, and nothing else has an author until they exist. Records start the same month so the first review has something to read.

Cadence and Signal come second because a rhythm needs a month of memory to read against, and drift needs a baseline to drift from. The baseline is encoded intent, not last month's behavior. A system can hold steady against its own history while drifting from the point.

The portfolio segment comes last because funding decisions made on fewer than two monthly reads are opinions with a calendar invite. Intent and Policy get no drafting workshop on purpose: written cold they would be one more design-time artifact, so they enter through the record template and are pressure-tested into shape by the first quarterly.

The blocks stack rather than hand off; after day ninety all six run together, permanently. And ninety days promises no harvest. It promises the smallest version of the loop that can run itself and be read. The harvest claim lives in section eight, on a 2028 clock, where it can be falsified.

Days 1 to 30: name the Roles. One person per floor-function with runtime authority made explicit. Stand up the record template: decision, basis, conditions, alternatives rejected, owner. Before the first record is written, recover the real process from the people running it. The stated process is never the running one, and you cannot encode what no one can say. Start writing Records at F4 and F1 first, where the cost of amnesia is highest.

Days 31 to 60: set the Cadence. The monthly operational read against a consistent baseline frame. Define the first Signals worth surfacing: the drift that matters, per capability, with thresholds owned at F2.

Days 61 to 90: run the first quarterly portfolio segment. Fund, keep, uplift, combine, deprecate, or subscribe, decided on evidence, recorded callably. Read the two gauges for the first time and let them be uncomfortable.

What not to do in the same ninety days: hire for it, reorganize for it, or buy a platform for it. The operating layer is operating discipline first.

The run becomes a control loop, older than the tools. The science is the loop. The art is the refrain.

Figure 7 shows the order of operations at inception, and Figure 8 the refrains that follow. A turn of the loop is not a re-run of inception. Inception places the first primitives. The turns that follow work them toward the full six, the three pairs from Section four. Then keep going: extend them, integrate them with the systems already in the building, absorb what an acquisition or a new regulation forces. The shape stays constant. The work inside it does not. Begin in one place, and let each turn carry the firm from what it could encode first to what it needs to run on.

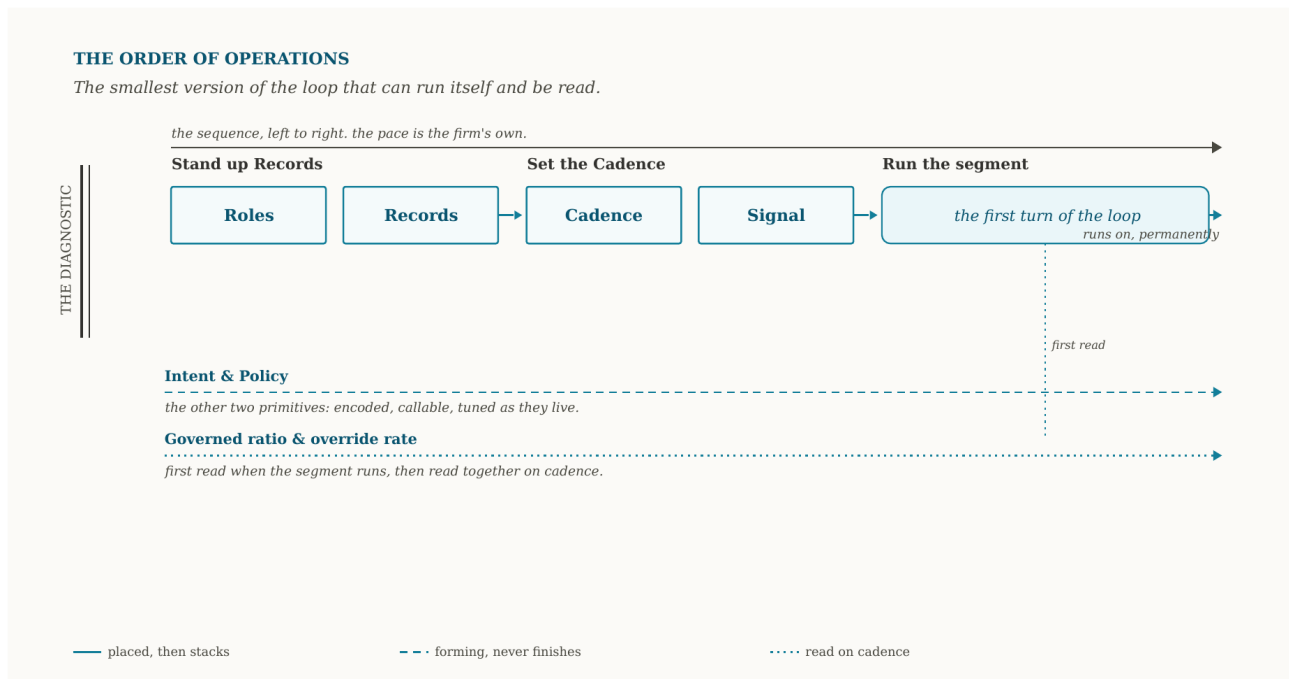


Figure 7. The order of operations. The sequence is fixed; the pace is the firm's own. Intent and Policy are encoded and made callable, then tuned as the firm runs.

ONE INCEPTION, MANY REFRAINS

Begin the same way. Then choose a rhythm, or write your own.

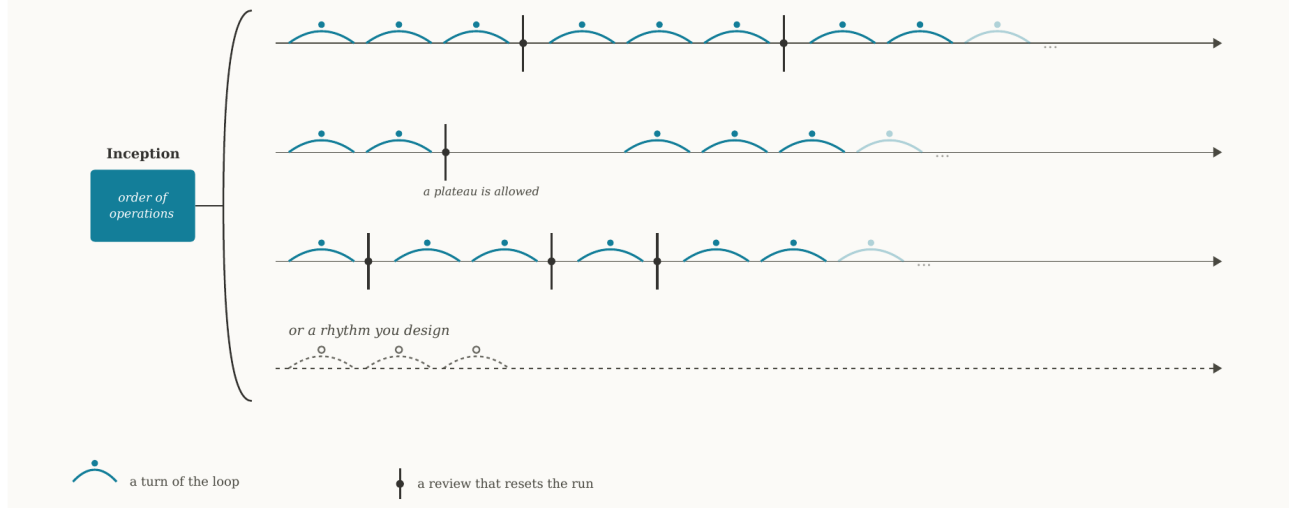


Figure 8. The refrains: design your own journey. Inception first; after that the rhythm is the firm's to play.

You can run the loop by hand at kickoff, on legacy and then prototype primitives, lean-startup style, and tighten it from there. Data is valuable but it is not the moat, and the carrier that runs the primitives is rentable, so it is not the moat either. The encoding is the moat: the firm's own intent, made callable. Host it and carry it however you like, or must, for compliance and sovereignty. Only make sure you own it.

Tools can carry it later; they cannot substitute for it.

The doing got cheap. The deciding did not.

The firms that thrive after the inversion will be the ones that give the deciding a floor to stand on, take the hand off the crank, and put the judgment where only judgment can go. That is what the ninety days begin to buy: a firm that stays coherent at machine pace, keeps its memory, and changes its own boundary on purpose. That's the lesson we take from Coase and Van Valen. And the opportunity at hand.

10. Lineage, credits, and what remains open

This framework synthesizes a practitioner's quarter-century in enterprise delivery with signals surfaced in public conversation, and it stands on named shoulders. Ronald Coase posed the boundary question in 1937. W. Edwards Deming assigned the system to management in 1986. Stafford Beer drew the viable system before either was fashionable. Erik Brynjolfsson and colleagues explained the J-curve trough. Shahidi, Rusak, Manning, Fradkin, and Horton posed the Coasean Singularity question; Howard Yu made it legible to operators. Eaneff, Obermeyer, and Butte named algorithmic stewardship for clinical AI; Perdomo, Zrnica, Mendler-Dünner, and Hardt formalized performative prediction, the mathematics under the behavior-change loop. Stefan Norrvald reached the structural diagnosis independently. Practitioners including Deepak Jha (the memory continuity layer and Judgment Capital Management), Julie Banfield (the Agent Decision Record), Maryna Deundiak (AI reveals what was never formally encoded), Akhilesh Warik (designed friction, cadence mismatch), and Ashwinikumar Patil (market forgiveness) sharpened specific load-bearing pieces, and are credited at the point of use in the essays this paper distills.

Open questions are flagged as open. Whether the six primitives are complete is a working claim, not a theorem. Whether *legitimacy* belongs in the set is unresolved. Whether the Coasean Singularity arrives in its strong form does not need to be settled for the inversion to bind today, and nothing in this paper waits on it.

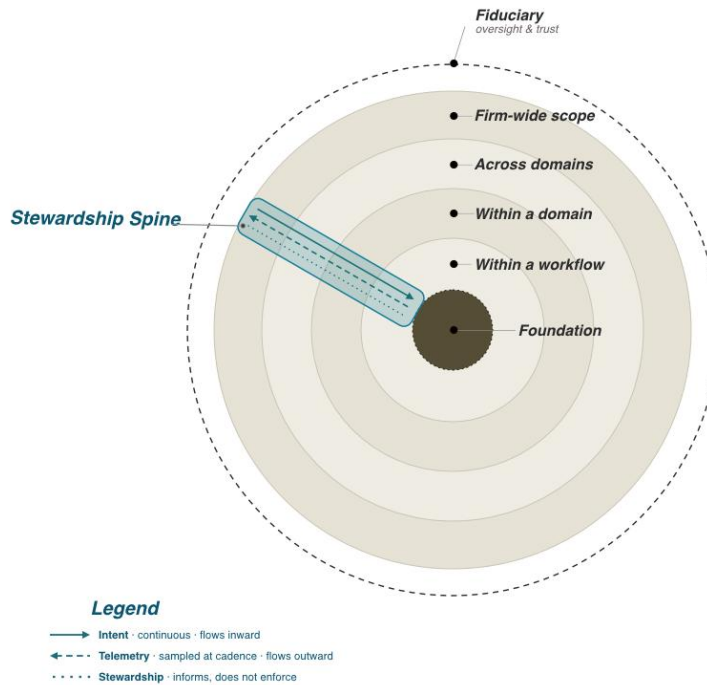
The essays behind this paper, with worked examples and the full argument, are published serially at Designed Friction (1aldean.substack.com). The patching account in section two is told at length in "The hand held it together, for a time" (May 2026).

Endnotes

About the author. Al Dean is a Cornell-trained chemical engineer, a delivery and systems-thinking leader with more than twenty years across enterprise strategy, operations, and program delivery, including building PMOs from scratch and recovering programs written off by others. Before enterprise technology he led a tank platoon and coordinated forward-deployed task force operations as a U.S. Army captain. He has held the loop this paper describes, by hand, and writes at Designed Friction to help us make sense, synthesize the way forward, and make a difference. <https://www.linkedin.com/in/al-dean/>

1. Gartner, "Gartner Forecasts Worldwide AI Spending to Grow 47% in 2026," press release, 19 May 2026. Worldwide AI spending forecast at 2.595 trillion dollars for 2026, up from 1.765 trillion in 2025.
2. McKinsey & Company, "The State of AI in 2025: Agents, innovation, and transformation," McKinsey Global Survey, 5 November 2025. 88 percent report regular AI use in at least one function; 39 percent report enterprise-level EBIT impact; no more than 10 percent have scaled agents in any single function.
3. MIT Project NANDA, "The GenAI Divide: State of AI in Business 2025," July 2025. Scope note: the 95 percent figure covers custom and embedded enterprise GenAI pilots and measurable P&L impact, from 52 executive interviews, 153 survey responses, and 300+ public deployments; the report describes its figures as directionally accurate. The same report finds externally partnered builds succeeding at roughly twice the rate of internal builds.
4. S&P Global Market Intelligence, 451 Research, Voice of the Enterprise: AI & Machine Learning, 2025, as reported by CIO Dive. Companies abandoning most of their AI initiatives rose to 42 percent in 2025 from 17 percent in 2024; the average organization scrapped 46 percent of AI proofs of concept before production.
5. Gartner, "Gartner Predicts Over 40% of Agentic AI Projects Will Be Canceled by End of 2027," press release, 25 June 2025. A forward-looking prediction, not a measured cancellation rate.
6. Stanford Institute for Human-Centered Artificial Intelligence, The 2026 AI Index Report, Economy chapter, April 2026.
7. Thomas H. Davenport and Randy Bean, MIT Sloan Management Review, 2026 executive survey coverage: 39 percent of Fortune 1000 firms report AI at production scale, up from under 5 percent two years prior.
8. PwC, 2026 AI Performance Study, surveying 1,217 executives across 25 sectors.
9. W. Edwards Deming, Out of the Crisis, MIT Press, 1986, p. 315. For contemporary re-measurements of the ratio, see the PwC split at note 8 and the BCG allocation at note 12.
10. Anthropic, "Project Glasswing: An initial update," 22 May 2026. anthropic.com/research/glasswing-initial-update.
11. Leigh Van Valen, "A New Evolutionary Law," Evolutionary Theory 1 (1973): 1–30, coining the Red Queen hypothesis after the racing scene in Lewis Carroll's Through the Looking-Glass (1871). The enterprise-strategy reading, that a firm must keep investing to hold relative position in a field where all rivals adapt, is now common in competitive-dynamics writing; this paper credits the biological origin and joins it to the Coasean inversion as the second of two independent pressures.
12. Boston Consulting Group, "AI Adoption in 2024: 74% of Companies Struggle to Achieve and Scale Value," press release, 24 October 2024. Survey of 1,000 CxOs across 59 markets; difficulty attributed roughly 70 percent to people and process, 20 percent to technology, 10 percent to algorithms. The lineup's local-optimization arithmetic is Eliyahu M. Goldratt's: The Goal, North River Press, 1984, and the Theory of Constraints it introduced.

13. Peyman Shahidi, Gili Rusak, Benjamin S. Manning, Andrey Fradkin, and John J. Horton, "The Coasean Singularity? Demand, Supply, and Market Design with AI Agents," in *The Economics of Transformative AI*, University of Chicago Press; NBER Working Paper 34468, 2025. The foundational boundary question is Ronald H. Coase, "The Nature of the Firm," *Economica* 4(16), 1937, pp. 386-405; its modern development is Oliver Williamson's transaction-cost economics, Nobel Prize in Economic Sciences, 2009.
14. Howard Yu, "Coase vs. Claude and the Future of the Firm," *One Inch Ahead*, 2 April 2026; and "The Coasean Singularity: Why AI Is Ending the Org Chart as We Know It," Inc., 5 May 2026.
15. Stefan Michel (IMD), LinkedIn post, December 2025, on the single-yardstick failure: when ROI is the universal metric, strategy collapses into short-term financial sorting; a core business, an emerging business, and an option each deserve their own question, and the latter two require strategic judgment rather than spreadsheet logic.
https://www.linkedin.com/posts/prof-stefan-michel_strategy-leadership-transformation-activity-7403418004501422080-FSAC. The underlying model is Mehrdad Baghai, Stephen Coley, and David White, *The Alchemy of Growth*, Perseus Books, 1999 (the Three Horizons).
16. Raja Parasuraman and Dietrich H. Manzey, "Complacency and Bias in Human Use of Automation: An Attentional Integration," *Human Factors* 52(3), 2010, pp. 381-410.
17. Sundar Pichai, Alphabet Q3 2024 earnings call, 29 October 2024 ("more than a quarter of all new code... then reviewed and accepted by engineers"); Satya Nadella at LlamaCon, 29 April 2025 (20 to 30 percent of code in company repositories, varying by language), as reported by TechCrunch.
18. Sebastian Siemiatkowski, Bloomberg interview, 8 May 2025. Klarna's move is best read as a re balancing toward a hybrid model, not a retreat from AI.
19. Stafford Beer, *Brain of the Firm*, 1972, and the Viable System Model; W. Ross Ashby, *An Introduction to Cybernetics*, 1956 (requisite variety); Stefan Norrval, "Axion: A Three-Axis Model of Organisational Coherence," Zenodo, 28 December 2025, DOI 10.5281/zenodo.18073324.
20. Stephanie Eaneff, Ziad Obermeyer, and Atul J. Butte, "The Case for Algorithmic Stewardship for Artificial Intelligence and Machine Learning Technologies," *JAMA* 324(14), 13 October 2020: 1397–1398.
21. Juan C. Perdomo, Tijana Zrnica, Celestine Mandler-Dünner, and Moritz Hardt, "Performative Prediction," *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020, PMLR 119: 7599–7609.
22. Council and Parliament provisional agreement on the Digital Omnibus, 7 May 2026, deferring EU AI Act Annex III high-risk obligations to 2 December 2027, pending formal adoption; California Privacy Protection Agency regulations on automated decisionmaking technology, finalized 2025, with obligations phasing in from 2027.
23. Erik Brynjolfsson, Daniel Rock, and Chad Syverson, "The Productivity J-Curve: How Intangibles Complement General Purpose Technologies," *American Economic Journal: Macroeconomics* 13(1), 2021, pp. 333-372.
24. See notes 6 and 7.



The Scope of Coordination

The firm seen from above.
One variable: *how wide coordination has to reach.*

The **center** is Foundation. AI primitives, networking, data, infrastructure. Every ring stands on it.

Move outward and each ring widens what must be coordinated. A single workflow. A domain. The whole firm. The outer ring is firm-wide scope, where intent is set.

The Spine is not a ring. It runs through all of them. Stewardship informs across every scope at once. It does not enforce.

People and AI work at every scope. What changes as you move out is not who acts. It is how far the coordination has to hold.

Exhibit: Coordination layers across scopes within the hybrid enterprise.