# The Enterprise AI Operating Model

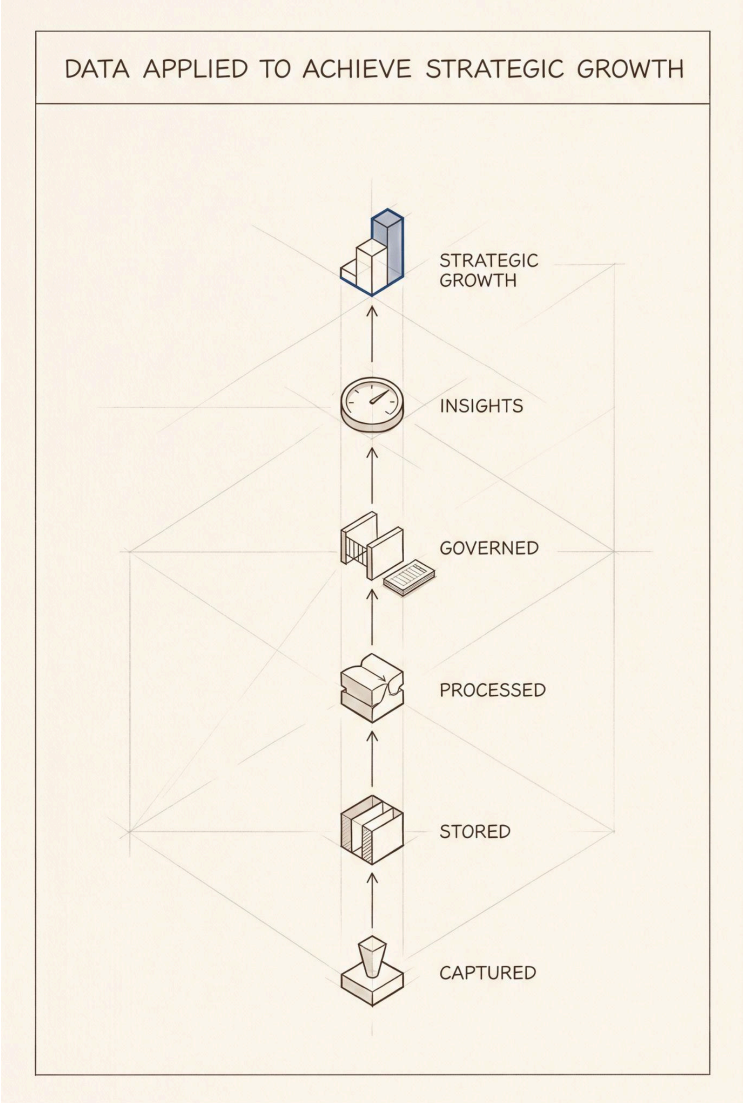
*Turning AI activity into value you can name, attribute, and defend.*

Publication Date 9/8/26

Version 1.0

Author Jeremy Taylor





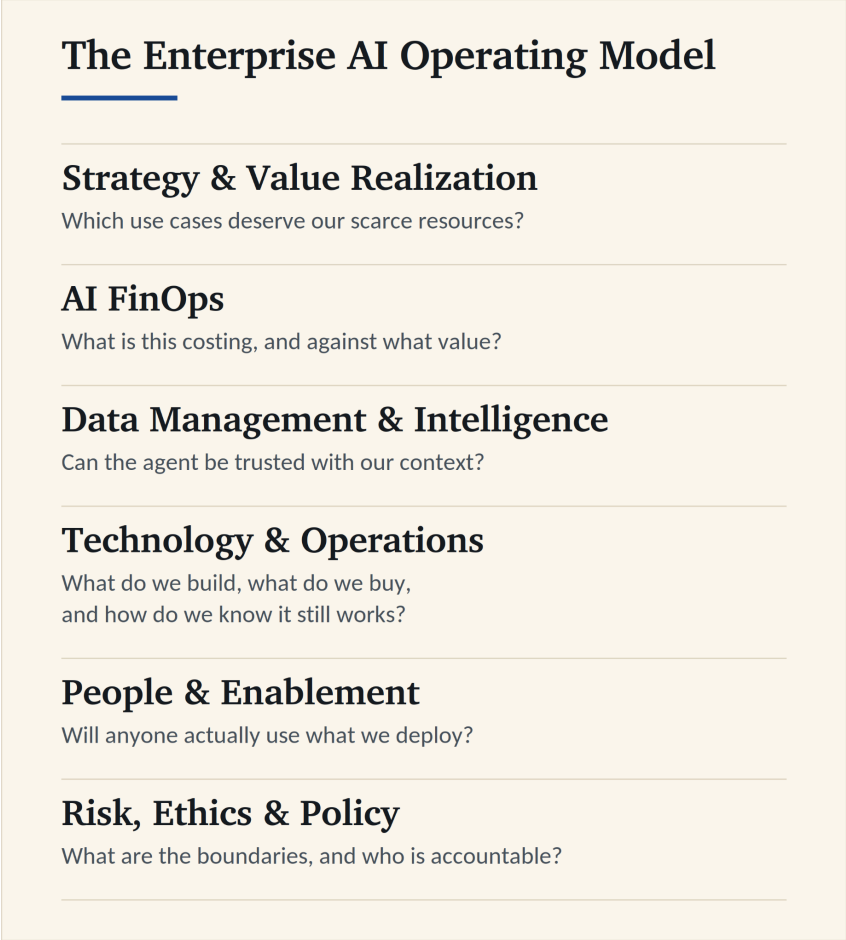
# Table of Contents

# The Enterprise AI Operating Model

Ask a leadership team what they're doing with AI, and you'll get a list. Ask which items on that list are paying for themselves, and the room gets quiet.

The gap isn't technology. Pilots run, licenses are procured, agents ship. What's missing is the operating discipline that turns activity into value someone can name, attribute, and defend.

How do we channel the frenetic energy around AI into a cohesive, governed purpose?



I created **The Enterprise AI Operating Model** as an artifact of personal experience and research while completing the Carnegie Mellon University Chief Data and AI Officer executive program (cohort 3). The six pillars are universal requirements for moving into the value phase.

## How to use this

This is for the person who owns the AI portfolio: the executive who has to say which projects continue, what they cost, and who answers for one that fails.

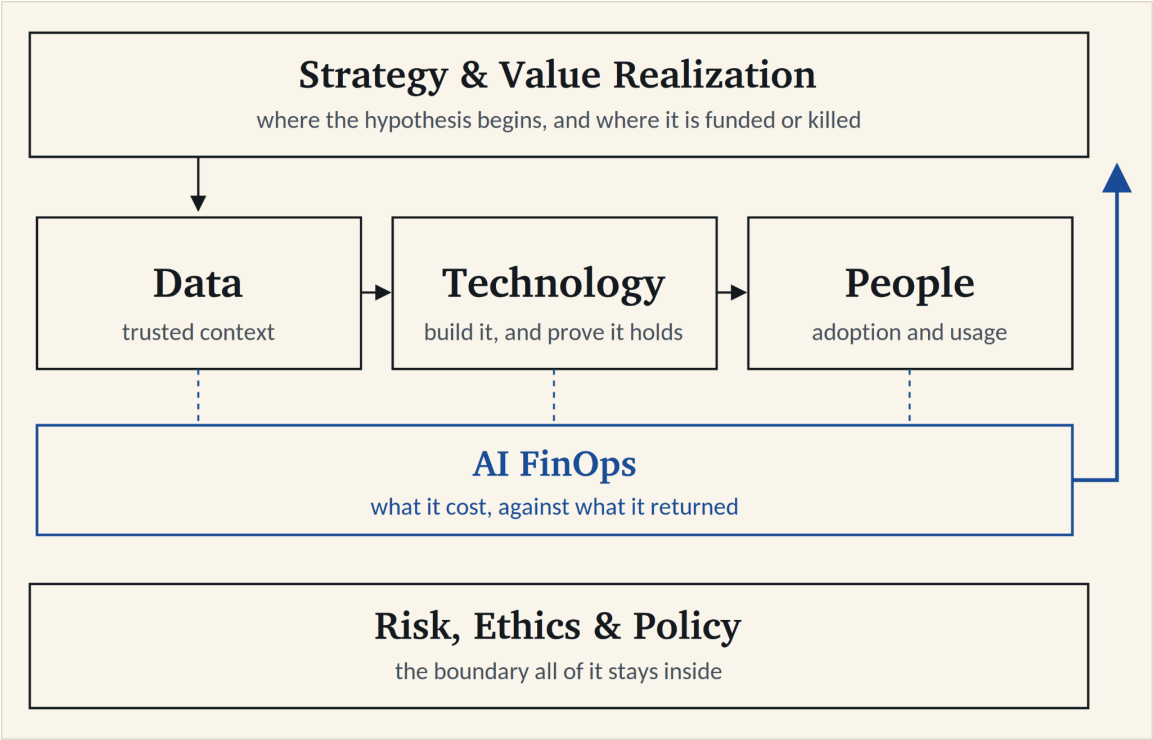
When I say AI in this document, I mean large language models and the agents built on them. Classical machine learning has its own cost structure, its own testing discipline, and its own governance questions.

This is a general model. It works at the level of how an enterprise organizes to get value from AI, and stops where the answer depends on your industry, your regulator, or your scale. Each of those deserves its own treatment.

## Where to Start

A hypothesis run through the other five pillars comes back funded or killed.

The first question is sequencing. In what order do we actually do this? Strategy is overarching and policy is the floor. The work happens in between, and it runs on a loop.




**Start with the two that cost almost nothing.** Strategy sets the value hypothesis. Policy sets the boundary. Neither needs a platform, a vendor, or a quarter. A hypothesis is a paragraph. A usage policy has two core rules.

These aren't gates that block everything behind them. They're the cheapest work in the model, and they make everything after them legible. Skip Strategy and you can't tell delivery from activity. Skip Policy and you haven't avoided the risk; you've only stopped seeing it. Your people are pasting company data somewhere today.

**Then build, in this order: Data, then Technology.** Most organizations invert this, and it's expensive.

A democratized platform is procured in weeks. A strategic application is built over quarters. Neither one can procure a trusted context layer, which is why Data comes first in both cases.

**Value shows up at People, or it doesn't show up.** Adoption is the last mile and the only one that pays. A deployed capability nobody opens returns nothing, however well the four pillars before it were executed.

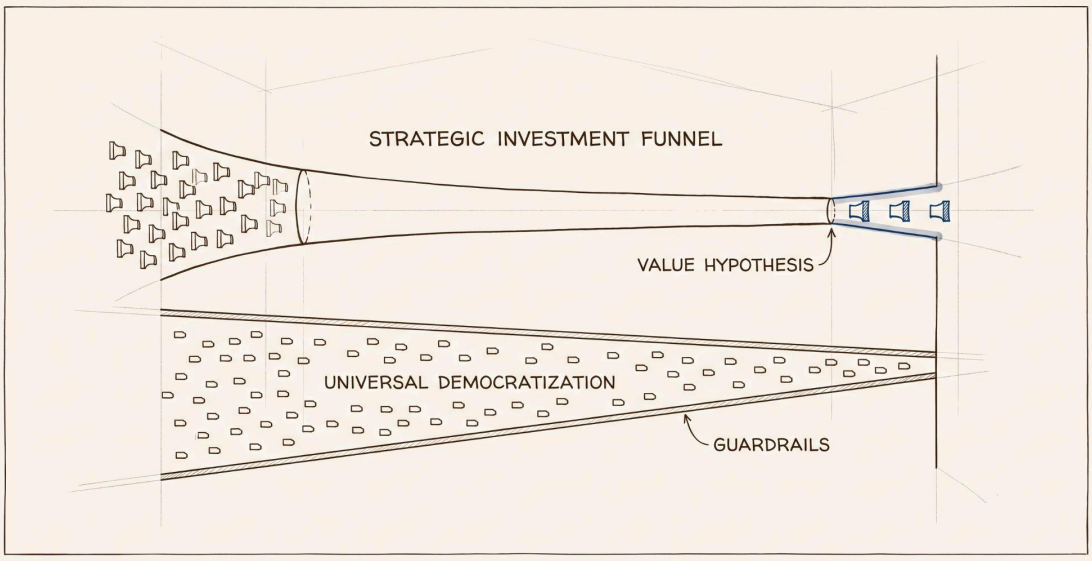
**AI FinOps measures all of it and reports back.** It runs underneath the work rather than at the end of it, which means it has to exist before the spend it explains. What FinOps returns to strategic review settles the hypothesis: what this cost, against what it produced.

Then the model loops. The hypothesis is proven or killed, unit cost is known, adoption is measured, and the next idea is written with better priors than the last.

# Pillar 1 - Strategy & Value Realization

Several teams are running pilots. Multiple leaders “own” AI platforms and are applying AI concepts to business domains. Collectively, executives see spend and experimentation but cannot articulate a cohesive approach.

Value Realization is the overarching organizing factor for the AI Operating Model. A successful AI strategy requires a split approach to capital and talent allocation.




Value follows ruthless rationalization of the funnel.

**The Strategic Investment Funnel** - This is for high-complexity, large-scale initiatives that carry unique risks. These projects require centralized governance and deep engineering rigor. Every project must begin with a Value Hypothesis. Measure actual performance against both leading and lagging KPIs, focusing not just on innovation but financially measurable cost avoidance and revenue growth.

**Universal Democratization** - To avoid becoming a bottleneck, the enterprise must enable self-service AI. Cloud provider platforms (AWS Bedrock, Azure AI Foundry, or Gemini Enterprise) allow teams to build their own capabilities within pre-defined enterprise safeguards. Integrating no-code AI tools is now a consideration for all enterprise positions (Pillar 5) and analogous to workforce adoption of spreadsheets. The risk is that without clear usage and deployment policies (Pillar 6), democratization becomes a significant source of unmanaged technical and compliance risk.

**An example value hypothesis, filled in**

A hypothesis that cannot fail is not a hypothesis. Six fields, and the last one is the one most organizations skip.

**Use case.** A B2B operations platform publishes a mature API surface, and customers already automate against it. But the work of noticing what matters in their data and deciding what to do about it happens outside the product in spreadsheets, in BI tools, in the heads of two or three power users per account. The proposal is an agentic layer that reads across the customer's own tenant data, surfaces the exceptions worth acting on, and executes a bounded set of actions through the APIs we already publish, under approval.

**Value hypothesis.** If customers can move from noticing to acting without leaving the product, accounts adopting the layer will expand rather than churn, because the layer displaces work currently carried by customer headcount or a services vendor.

**Leading KPI - 90 days.** Share of enrolled accounts executing at least one agent-initiated action per week, and the ratio of surfaced recommendations to accepted ones. Depth of use, not seat count: a seat that never acts proves nothing (Pillar 5).

**Lagging KPI - four quarters.** Net revenue retention in enrolled accounts against a matched control cohort.

**Unit economics.** Fully loaded inference and infrastructure cost per accepted action, tagged to this initiative from the first sprint (Pillar 2), held beneath a defined ceiling as

a share of incremental revenue per account. A cost you cannot attribute is a return you cannot claim.

**Kill criterion.** Two arms, because there are two ways to be wrong. If acceptance rate on surfaced recommendations sits below 30% at 90 days, the agent is not finding what matters, and no amount of interface work fixes it; stop rather than re-scope. If acceptance is healthy but retention shows no separation from the control cohort by the fourth quarter, the layer is valuable to users and not to the business. Reprice it or stop.

**Accountability.** A named business sponsor who accepts or rejects the claim, not the engineering lead who built it.

The discipline is writing the kill criterion while it's cheap to be honest.

How to **organize these tracks**? In practice, an AI Center of Excellence wrapper is an efficient organizational approach. Bring in the leads for strategy, finance, engineering, and legal. Set a cadence to manage the committee decisions, funnel review, and KPI review, and the enterprise achieves ***managed*** AI. A hybrid of **centralized standardization with federated business unit engagement** works well**.** Without standards, scale will flounder. Centralize the core business rules for how AI will get built. But, the business units have the knowledge to sign off on the value hypothesis for any project. Their expertise is required in refining requirements, success criteria, and establishing the KPIs to measure outcomes. Business teams drive, sponsor, and accept that the work achieved the original value claim.

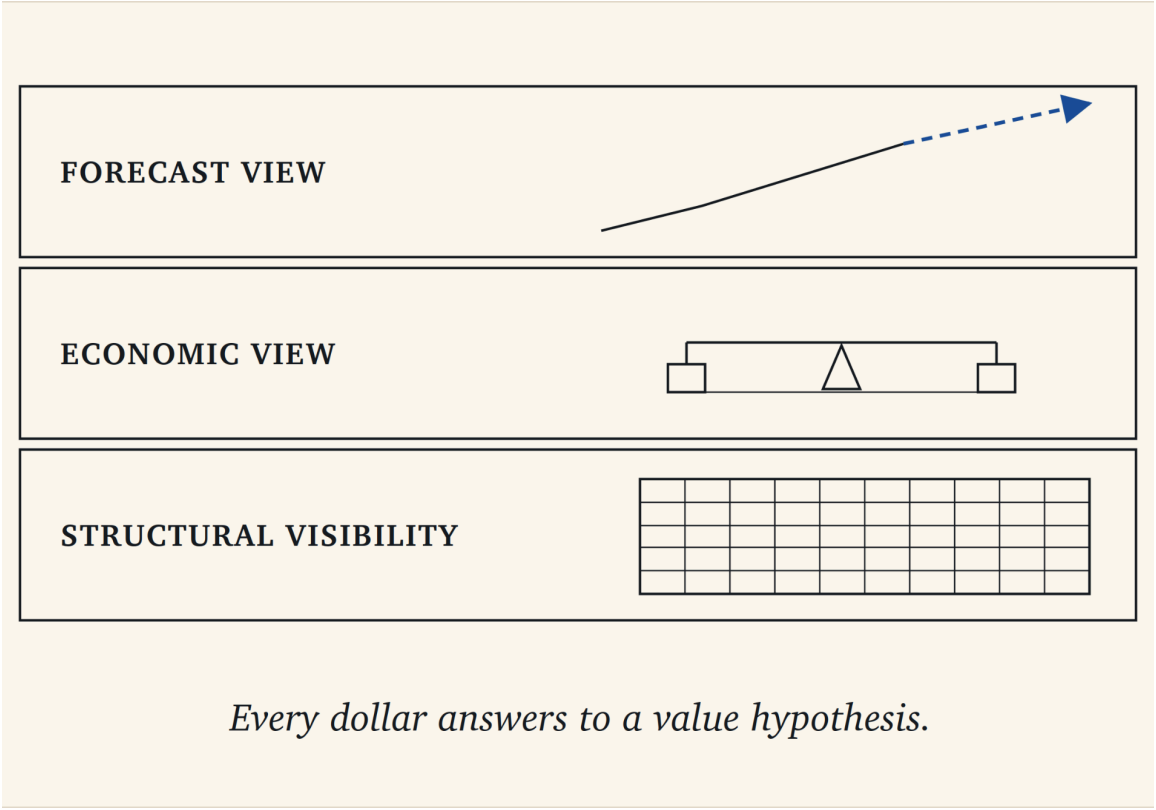
*Value follows ruthless rationalization of your investment funnel.*

# Pillar 2 - AI FinOps

The invoice arrives, and it's larger than anyone expected. Nobody in the room can say which initiative caused it, which means nobody can say whether it was worth it. Somewhere inside that number is one model stuck in a loop. Somewhere else inside it is the only project actually paying for itself. At this level of detail, they look identical.

AI FinOps solves this problem, and it's a prerequisite for ROI rather than a stage of maturity. Strategy and Value Realization tells you what to cut, but FinOps tells you what each option costs, by correlating spend against the value hypothesis.

To keep this simple, let’s make the assumption of using off-the-shelf LLMs (e.g., Gemini, Opus) in one of the major cloud providers. A simple monthly aggregate view is insufficient. You must be able to correlate costs to initiatives to measure ROI.



## Stage 1 - Structural Visibility

Within the cloud provider's standard billing reports, create two distinct views to map your environment:

**The Org-Wide SKU View**: A complete visualization of all specialized LLM AI SKUs and inference costs across all projects. This is critical for identifying unmanaged decentralized consumption and exposing hidden costs. Set budget thresholds to avoid a catastrophe.

**The Application Cost View**: DevOps teams must rigorously label/tag every infrastructure component associated with prioritized, high-stakes AI funnel projects. This provides a comprehensive picture of the full cost including Kubernetes clusters, databases, and networking.

## Stage 2 - Economic View

Visibility is only the ***foundation***. As your operating model matures, transition the aggregate data into economic models:

**Unit Cost Methodology -** Move from "monthly total" to a defined unit cost per transaction or token.

**Chargeback Methodology -** Develop a mechanism where these unit costs are charged back to the specific business units consuming the resources. Accountability optimizes consumption.

Whether looking at unit cost or chargeback, be realistic in supporting adoption timelines (Pillar 5) feasible for value to emerge.

## Stage 3 - Forecast View

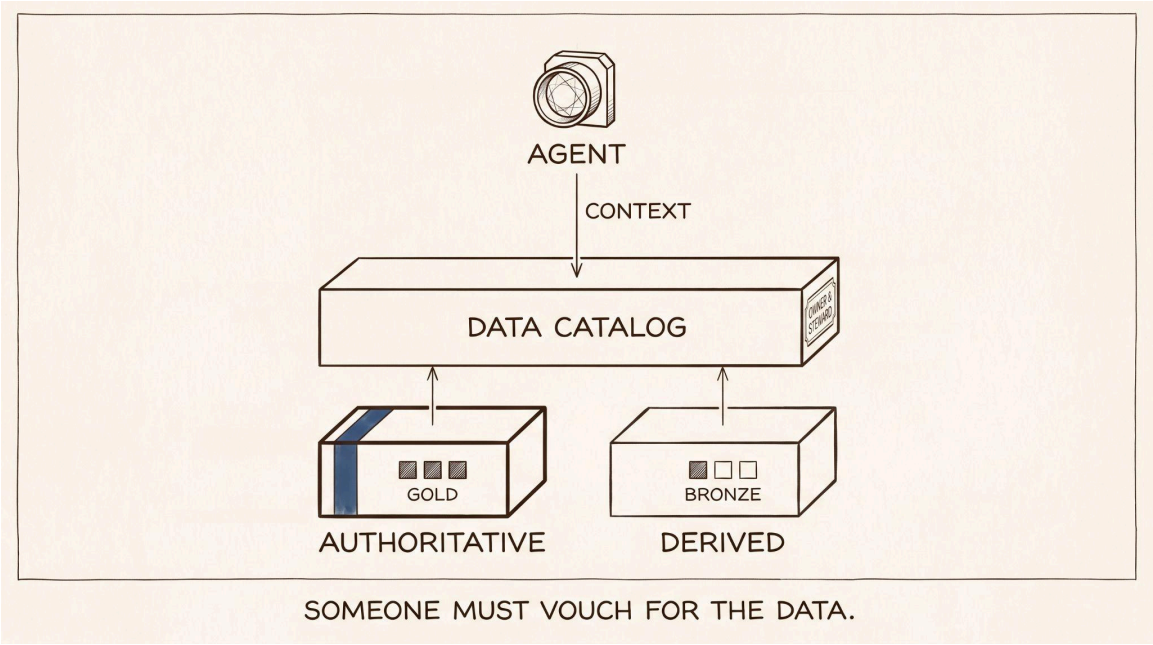
Finally, true AI FinOps maturity combines anchored historical, allocated financial data with a forecast. Accurate forecasting depends on locking down assumptions across several variables:

- The volume of production versus experimentation use cases in the pipeline.
- Projected audience size and adoption/retention rates for each use case.
- The expected mix of advanced versus fast (low-latency) model calls.
- The addition of ancillary infrastructure costs (memorystores, Vector DBs).

*Every dollar answers to a value hypothesis.*

# Pillar 3 - Data Management & Intelligence

Every leadership team knows this bottleneck: executives need ad hoc analysis, requests pile onto the BI team, and speed-to-decision suffers. Now picture an agent that returns the analysis faster than a request could even reach the queue, leaders self-serving, and load coming off the BI team. That's achievable today, and the pattern behind it is tested and production-ready.




None of it works without the data layer underneath. LLMs are only as good as the context they're given, and Data Management is how you supply it. An agent's answer is only as trustworthy as the source it referenced, and the agent cannot judge that on its own. Something has to tell it. That is the whole job of the data layer, and it reduces to three questions: what does this mean, who says so, and how much should I trust it.

A **Data Catalog** answers all three. It builds on Data Mesh and Data Product concepts (a mesh isn't required, but it scales this well), and it is where ownership and stewardship live, exposed to agents over MCP.

**Meaning.** Fields, business terms, and query examples, including the logic a subject-matter expert applies without thinking. The unwritten rules everyone in the department knows but have not documented are the technical debt that must be paid.

**Ownership.** A named person must be accountable for the product, not a committee. Accountability is what makes an answer defensible when someone challenges it, and an answer no one will defend is not decision-grade.

**Grading.** Identify one product as authoritative for each business concept so the agent knows which of several similar sources to use. Grade the maturity separately (gold, silver, bronze). Authority and maturity are different questions: the authoritative source for a concept may still be bronze, and an agent that understands this is more trustworthy.

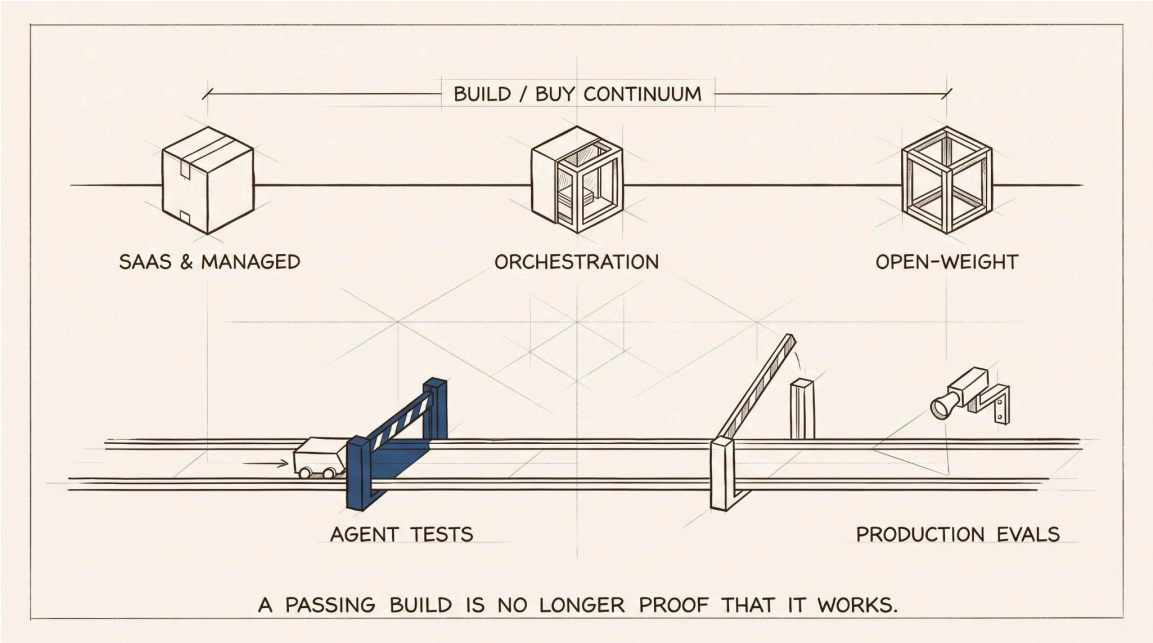
Start with one dataset that is well understood, high quality, and matters to the business. Make it a data product, expose it, and let the pattern prove itself before you scale. Pair it with a minimally invasive governance program and you scale confidently instead of guessing.

> *Someone must vouch for the data.*

# Pillar 4 - Technology & Operations

Consider the system operator who knows the software test suite passed for this deployment but is watching AI agent performance degrade. New technology requires new tools and methods for enterprises to do business.

One factor is new; one decision is familiar. Probabilistic solutions change the rules for testing and operations. And, every enterprise still has to decide what it builds and what it buys.




**The Build vs. Buy Continuum** - All hyperscaler cloud providers offer ecosystems, ranging from no-code LLM agents to custom code integrated with a model of your choice. Your platform decision is likely driven by your incumbent cloud provider and the complexity of the application.

- SaaS & Managed Platforms: Specialized AI products accelerate the journey but force typical trade-offs in portability, flexibility, and cost.

- Custom Orchestration & Harnesses: Using frameworks like the open source LangChain or cloud developer suites like Google Vertex AI allow teams to build highly customized application logic, manage state, and orchestrate complex workflows. This delivers maximum flexibility with a steeper learning curve.

- Open-Weight Architecture: Highly technical teams might even consider deploying community-licensed or open-weight models (like Qwen) on purchased compute. This provides distinct operating cost predictability (Pillar 2 - AI FinOps) where you may consume as many tokens as your purchased compute delivers.

***Strategic funnel work deserves a build***. If the application contributes to competitive advantage, then it requires the flexibility and portability without the hindrance of a vendor roadmap limitation. Give yourself the option of full optimization.

**The Probabilistic SDLC** - Quality is always a concern in software, but moving from deterministic logic to probabilistic models changes the game. The standard protections of a Software Development Life Cycle (SDLC) must evolve to manage AI Risks (Pillar 6).

- Agent Test Frameworks: Teams must identify repeatable test scenarios for continuous regression testing whenever a prompt, model, or context window change is introduced. Automate this as a gated check in your release process.

- Production Validation: Post-deployment, using an LLM as a judge provides crucial automated evals regarding agent performance and user experience.

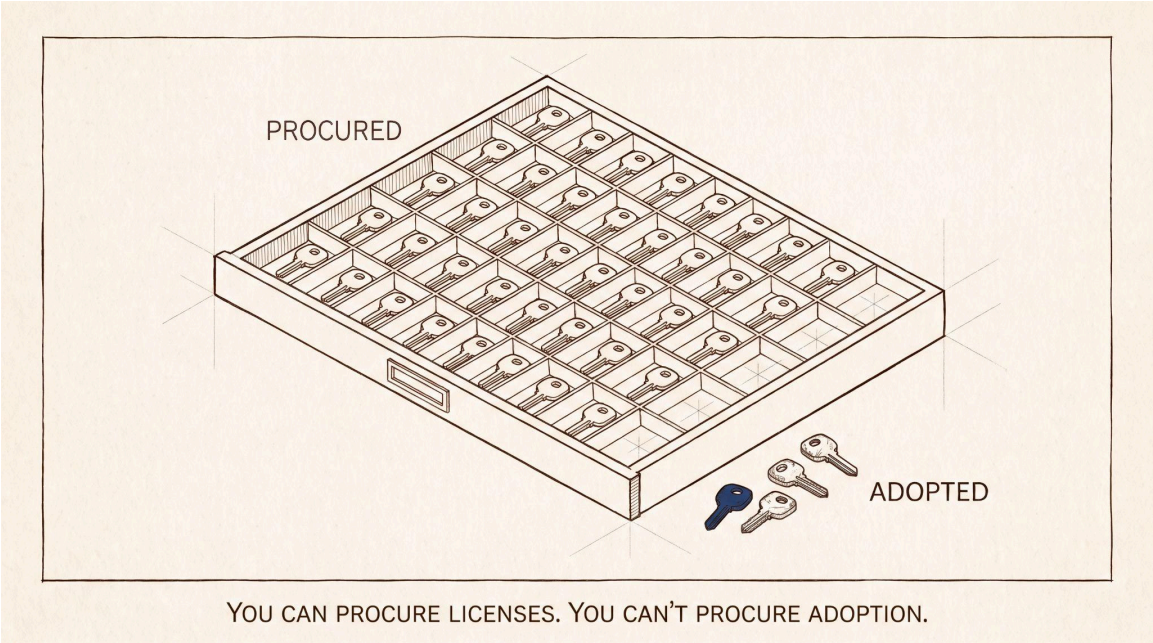
*A passing build is no longer proof that it works.*

# Pillar 5 - People & Enablement

Six hours. That's the annual efficiency gain a $120k knowledge worker needs to break even on a $30/month license. About 90 seconds a workday.

Which means the democratization buy decision is not the hard part. If breakeven is 90 seconds a day, seat unit cost (Pillar 2 - AI FinOps) isn't your risk; the seat nobody opens is. Unused licenses are all cost and no lift, and no procurement win rescues a 40% activation rate.

So Pillar 5 isn't a licensing question. It's an adoption question, and it breaks into three.




**Access - who gets a seat.** Some roles need no pilot: engineers, support teams, marketing and content, technical writers, analysts. High-volume text and code, measurable output, evidence-backed. Everywhere else, pilot before you scale and let team leads prove the value before you commit broadly.

**Capability - training has to be differentiated.** One generic prompting session serves no one. Executives need fluency in the art of the possible. Engineers need agents, harnesses, and evaluations. Business users need prompting and, more importantly, where the technology fails. Three audiences, three curricula.

**Belief - communication is the adoption lever.** People don't underuse AI because it's hard. They underuse it because they're unsure what's in bounds or unsure what it means for their job. Name your change agents, make experimentation explicitly safe within policy, and give people a stage: showcases, hackathons, and internal wins.

One nuance: studies often show the biggest lift is among novices. My experience is that experts surprise me with creative uses I didn't anticipate.

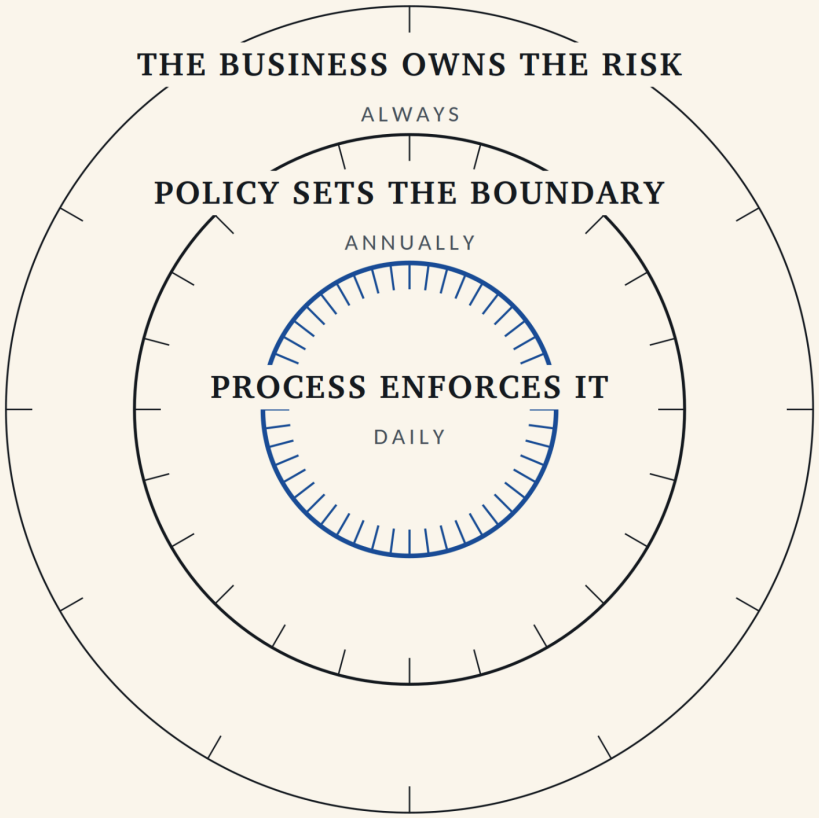
Enablement without guardrails is simply faster exposure. That's next in Pillar 6.

> *You can procure licenses. You can't procure adoption.*

# Pillar 6 - Risk, Ethics, & Policy

Your employees are already using AI. The question is whose terms govern the data pasted in and the applications consuming that data.

Whether or not you've deployed tools, people are solving problems with whatever they can reach. And, if you have deployed a tool, some of them still prefer a favorite. Pillar 5 was about enabling people. Policy is the counterweight.




*Governance isn't the tax on AI value.*
*It's the reason anyone trusts the answer.*

**Usage Policy** - Two rules do most of the work. First, company data only goes into enterprise-deployed tools where a vendor agreement governs retention and training. That one rule puts most of your exposure under contract. Second, define classifications appropriate to the organization (Public, Internal, Confidential, Restricted) and put ***Restricted*** permanently out of bounds. Then say plainly what ***is*** allowed. A policy that only prohibits is a policy people work around.

**Deployment Policy** - Three risks lead: financial (unbounded costs), autonomy (which writes and actions require human approval), and access (whose privileges the agent is actually using). The third is the one that gets missed. Most agents are built to run under a service account because that's what makes them work reliably. But a service account has to see everything any user might ask about, which makes the agent the one identity in your environment that can read across every boundary the underlying systems enforce. Agents should act under the requesting user's identity.

Teams shipping without oversight aren't being careless. They're moving at the speed the business requested, which means oversight must be fast enough to keep up.

**Process** - Example questions to answer before anything ships:

- Should this use case exist? An agent making hire/fire decisions is a bad fit, for example.
- Is the data trusted and is it permitted as context?
- How will the application be observed and monitored?
- Who is accountable for the outcome?

The Data Catalog from Pillar 3 is where classification stops being a PDF and becomes enforceable metadata.

The Business owns the risk. Policy sets the boundary. Process enforces it daily, minimally invasive or it won't survive a deadline.

> *Governance isn't the tax on AI value. It's the reason anyone trusts the answer.*

# What Changes

Implementing the Enterprise AI Operating Model does not require a transformation program. Organizations can start where they are today. Every pillar can be started by one person in one quarter at a pace the business can support. The data catalog is a core component that may be new, but a catalog creates alignment and data product thinking. Move at whatever scale the business can support.

What you're building toward isn't a maturity score. It's a change in what the conversation sounds like. Today the question is "what are we doing with AI," and the answer is a list of activities. When the operating model is working, the question is "which hypotheses are still alive, what did they cost, and who is accountable for the next one?" And, there are answers.

AI doesn't become an advantage because you deploy more of it. It becomes an advantage when the organization can tell the difference between the projects that are working and the ones that are merely running.

# Continuing the Conversation

If you're working through any of this or think part of the model is wrong, I'd like to hear about it: jeremy.taylor@ossnetworks.com or linkedin.com/in/jeremytaylor00. Additional articles are at ossnetworks.com.

# About the Author

I lead global data and AI organizations with twenty years in technology. I've spent the past eleven years building and transforming cloud data platforms. Most recently, I've focused on adding an AI platform layer and agents in production.

I completed the Carnegie Mellon University Chief Data and AI Officer executive certificate in 2026. This model came out of that work.

# Sources and Influences

This document is a synthesis. Where it builds on other people's work, that work is named here. Everything not named below is my own.

**Governance.** The "minimally invasive" framing in Pillars 3 and 6 echoes Robert S. Seiner's Non-Invasive Data Governance™ (Technics Publications, 2014), a registered trademark of Seiner and KIK Consulting, used here as an adaptation rather than a restatement.

**Data architecture.** The data product and ownership concepts in Pillar 3 follow Zhamak Dehghani, *Data Mesh: Delivering Data-Driven Value at Scale* (O'Reilly, 2022). This model does not require a mesh implementation, but it borrows the product and ownership framing.

**Catalog conventions.** Authoritative-source designation and gold/silver/bronze maturity grading are common conventions across enterprise data catalog products rather than any single author's framework. Note that gold/silver/bronze here denotes *product maturity* and is distinct from the bronze/silver/gold medallion architecture used elsewhere to describe pipeline processing stages.

**Cost management.** AI FinOps in Pillar 2 extends the cloud financial management discipline codified by the FinOps Foundation. The three-stage progression (structural visibility, economic view, forecast view) is this author's adaptation of that discipline to AI workloads.

**Agent context.** The Model Context Protocol (MCP) referenced in Pillar 3 is an open standard published by Anthropic.