CONCIERTO | SOVEREIGN INTELLIGENCE

Sovereign Perspectives · Economics of Owned Intelligence

AI & Financial Determinism

How Predictable, Owned Cost Lifts the Ceiling on Ambition

Series Sovereign Perspectives
Category The Economics of Owned Intelligence

Format Essay
Published by Concierto



Version 1.0 · September 2026

Copyright © 2026 Concierto. All rights reserved.

Contents

Executive Summary

1. The Meter Sets the Ceiling
2. What Scale Actually Delivers
3. An Illustrative Model: Acme Enterprises
4. Lock-In and Escalation
5. Rationing Creativity
6. Predictable Cost Frees Ambition
7. What Leaders Must Do
8. The Choice

About This Series · Sources

Executive Summary

A cost you cannot predict is a cost you learn to fear, and a team that fears its AI bill quietly stops reaching for what AI could do.

When intelligence is rented and billed by consumption, its cost is unpredictable by design, and the response inside an organization is silent and corrosive: teams ration their own ambition, and their own headcount on AI, to stay under a budget they cannot forecast. This is the hidden tax of the metered model, larger than the bill itself, because it is paid in the work not attempted and the people not empowered. The companies capturing the largest returns from AI are the ones deploying it broadly and continuously, and broad, continuous deployment is precisely what an unpredictable meter deters. This paper argues that predictable, owned cost is not merely cheaper; it decides how many of a nation's or a company's people get to work with AI at all, and it draws on public company results and an illustrative five-year model to show both the size of the prize and the trap that keeps most organizations from claiming it.

In Brief

- Consumption-priced AI makes cost unpredictable, and unpredictable cost makes organizations cautious in ways that never appear on a budget line.
- To avoid catastrophic token bills, the rented model rations AI to a small tier of power users, so most of the workforce is left off AI entirely.
- A locked-in pricing model escalates cost over time, and consumption growth produces a violent cost divergence, the token trap, a few years in.
- Predictable, owned cost puts the whole workforce on AI and holds the cost to a controlled, predictable climb as usage grows, putting the whole workforce on AI rather than a rationed few.
- The real return on ownership is not only a lower bill; it is a higher ambition and a far larger share of the workforce empowered to create.

The deepest cost of a meter is not the money it takes. It is the ambition it deters and the people it leaves off AI entirely.

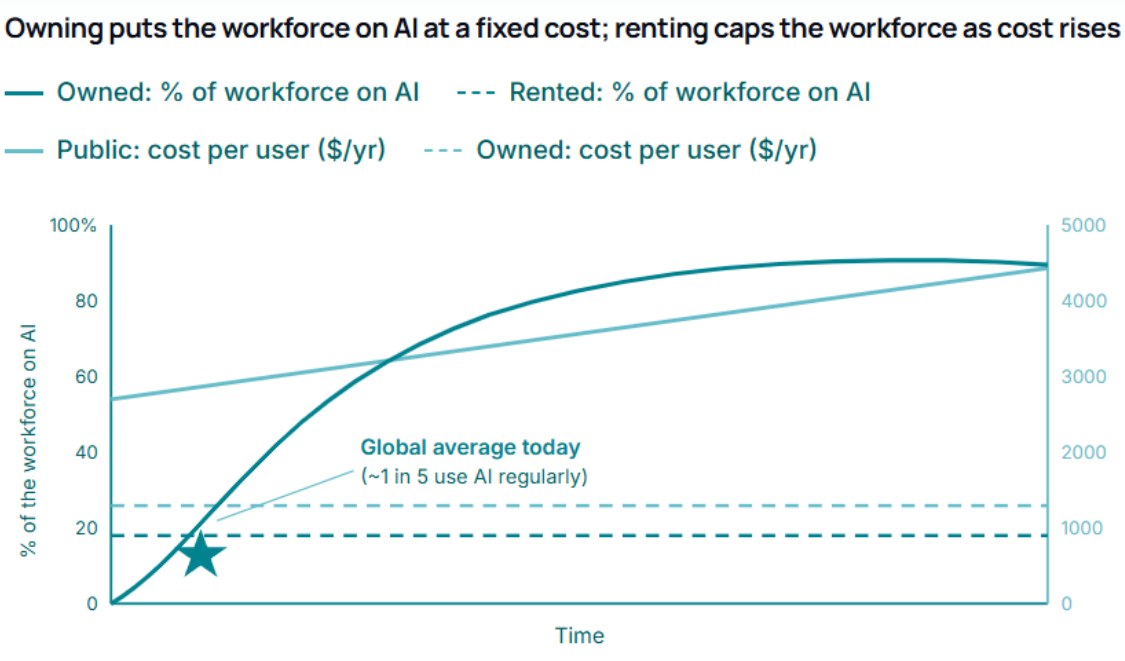
Copyright © 2026 Concierto. All rights reserved.

1. The Meter Sets the Ceiling

Give an organization a capability and an unpredictable bill, and it does not deploy that capability to everyone. It rations it to the few whose usage it can justify.
An unpredictable cost changes behavior long before it changes a budget, and its clearest effect is on who is allowed to use AI at all. Faced with a bill that rises with every user and every query, and that no one can forecast, an organization does the sensible thing: it restricts access to a small group of power users, and leaves the rest of the workforce off AI to keep the meter under control.

The ceiling is not set by the mission or by the technology's potential; it is set by the fear of the bill. And because that ceiling is invisible, no one debates it.
The organization simply decides, a thousand cautious decisions at a time, that most of its people will not work with the defining capability of the age.

An unpredictable cost does not just cap a budget. It caps how many of your people are allowed to think with AI at all.



ILLUSTRATIVE. Dark teal = share of workforce on AI (left); lighter teal = cost per user (right). Public rises; owned is fixed. Today marker: Gallup / U.S. Fed. 2025.

Figure 1. Share of the workforce on AI (navy, left) and annual cost (orange, right), owned versus rented, with today's global average marked. Illustrative.

Copyright © 2026 Concierto. All rights reserved.

2. What Scale Actually Delivers

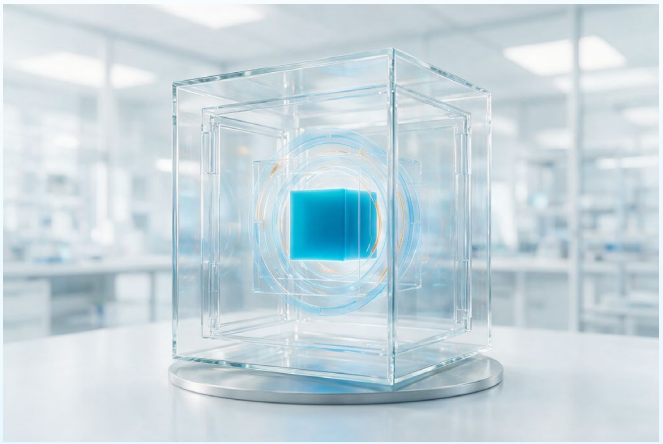
The prize for putting AI to work broadly is not theoretical. Companies operating it at scale have reported it in public, and the numbers are large.

Consider what organizations that have pushed AI deep into their operations have said publicly. The pattern is consistent: the returns come not from a single clever pilot but from breadth and continuity, from putting AI to work across many functions and running it at scale. These are company-reported figures, and they should be read as claims a company makes about itself rather than audited facts, but taken together they mark the size of what broad deployment can deliver.

Organization	What it deployed at scale	Reported result
Klarna (payments)	An AI customer-service assistant across its support operation	Work equivalent to about 853 full-time agents and roughly US$60M in annual savings by 2025; resolution times cut by about 82% (company-reported)
JPMorganChase (banking)	Hundreds of AI use cases in production and an AI assistant across its workforce	More than 450 use cases running in production; a contract-review system reported to reclaim roughly 360,000 lawyer-hours a year

Klarna's story carries the most instructive detail, because it did not end at the headline. Having scaled its assistant aggressively, the company then deliberately rebalanced, bringing human agents back into specific flows after concluding that cost-driven automation had, in places, produced lower quality, and committing that customers could always reach a person. The lesson is not that scale failed; the same company reported its assistant doing more work and saving more money a year later. The lesson is that the organizations which capture and keep the returns are the ones that can deploy broadly and govern quality at the same time, which is far easier when the intelligence is owned and tunable than when it is rented and fixed.

The returns from AI belong to those who deploy it broadly and govern it closely. Both are far easier when you own it.



Copyright © 2026 Concierto. All rights reserved.

3. An Illustrative Model: Acme Enterprises

To see how the ceiling translates into money and into people, follow one illustrative enterprise across three years.

Acme Enterprises is a large organization facing an AI disruption window with a complex legacy estate: roughly 3,000 users, a mix of knowledge workers and software engineers, and some 200 core applications that need modernization. Under a traditional cloud-first mandate, Acme would migrate and modernize that estate before deploying production AI, years and millions in technical cost before a single business outcome. The alternative inverts the order. By deploying an owned, sovereign AI platform on its existing infrastructure, Acme extracts AI value on day one, empowers its workforce immediately, and uses the resulting productivity to fund the technical migration that follows.

The cost comparison turns on one fact: the public cost is a moving target that rises every year, while the owned cost is fixed. On people, public intelligence averages about $2,700 per user per year today and climbs; the owned platform runs about $1,300 per user and holds. But the decisive line is agents. Acme runs a growing fleet of them, on the order of two hundred at this size, and an agent runs heavy and around the clock, so on a public model it costs close to a heavy engineer, roughly $9,420 a year, against about $2,000 owned, a gap of nearly five times per agent that widens as the fleet grows. Put people and agents together and, all in, Acme saves about 49 percent in the first year, widening to about 69 percent by year three as the public bill compounds and the owned one does not. The human figures match Acme's own measured cost of about $2,500 per user, so the comparison rests on the company's real data.

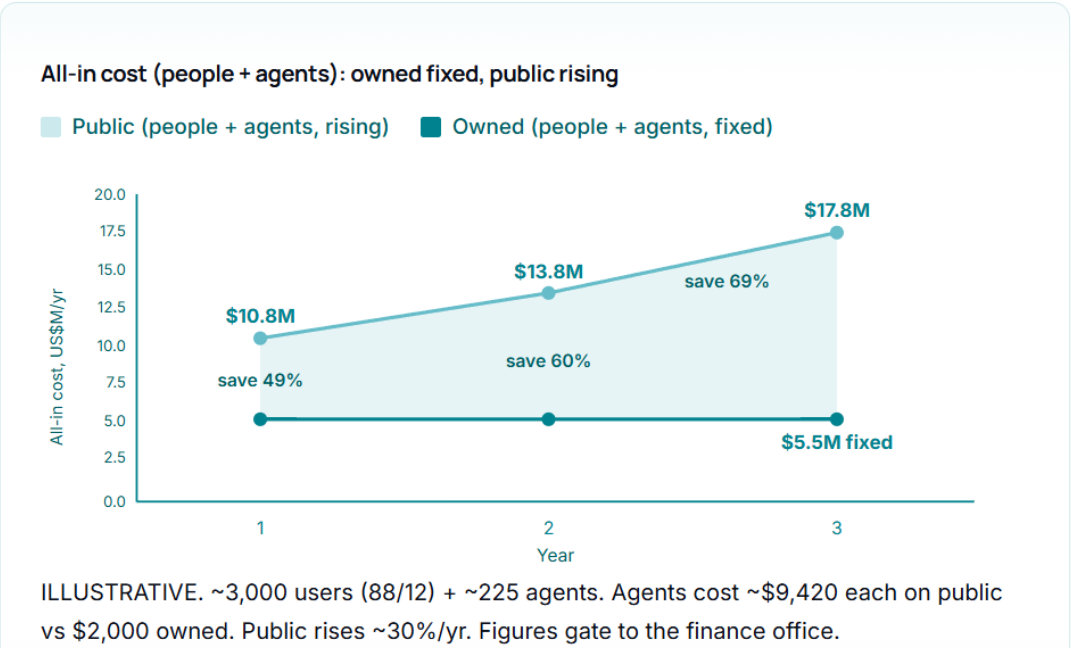


Figure 2. All-in annual cost, people and agents together, owned (fixed) versus public (rising). Illustrative; figures gate to the finance office.

Copyright © 2026 Concierto. All rights reserved.

Who gets empowered follows directly. A rising, unpredictable cost forces an organization to ration AI, both the people who get a seat and the agents it can afford to run. A fixed, lower cost does the opposite: it lets Acme put its whole workforce on AI and field a full fleet of agents at a price finance can forecast. Because the owned cost holds while the public one climbs, the same budget reaches more of the digital workforce every year, about twice as much in year one and close to three times by year three. The agent line drives most of that widening, because it is the fastest-growing and most cost-divergent part of the estate.

Acme, ~3,000 users + ~225 agents	Public (rising)	Owned (fixed)
People, cost per user / yr	~$2,700	~$1,300
Agents, cost per agent / yr	~$9,420	~$2,000
All-in cost, Year 1	~$10.8M	~$5.5M
Customer saves, Year 1	reference	~49%
Customer saves, Year 3	reference	~69%
Reach for a fixed budget, by Year 3	a rationed few	~3x the digital workforce

Read the table beside the curve and the point of financial determinism is complete. The owned platform costs less on day one and, because it is fixed while the public average compounds, its advantage widens every year, funding a broader deployment rather than a narrower one. The saving is real and large; the larger prize is that a fixed, predictable, lower cost is what lets the whole workforce onto AI at all.

A rising cost rations AI to a few; a fixed, lower cost puts the whole workforce on it. That is the difference determinism buys.

Copyright © 2026 Concierto. All rights reserved.

4. Lock-In and Escalation

The unpredictability and the divergence are not flaws of a young market. They are a position of strength for whoever sets the meter.

A pricing model the buyer cannot forecast, attached to a capability the buyer cannot easily leave, hands the supplier the power to escalate cost over time and leaves the buyer little recourse. The more essential the capability becomes, the weaker the buyer's position, because the cost of leaving rises with dependence, and a supplier that understands this can raise prices against a customer with nowhere to go. The token trap in the Acme model is not an accident of that customer's usage; it is the shape the rented model produces for every customer whose adoption succeeds. Lock-in and unpredictable pricing are one mechanism, and it compounds in the supplier's favor precisely as the customer comes to rely on it.

A price you cannot predict, on a capability you cannot leave, is not a purchase. It is leverage over you, and it tightens as you depend on it.

5. Rationing Creativity

The corrosive effect lands hardest on exactly the work that matters most, and on exactly the people most organizations most need to empower.

A predictable, broadly available capability invites experiment; an unpredictable, rationed one invites restraint. When AI is confined to the few who can justify a rising, metered cost, the rest of the workforce never learns to build with it, and even the empowered few build cautiously, avoiding the open-ended exploration that runs up the bill. Rationing creativity is the rational response to an unpredictable meter, and it is the surest way to waste the most creative technology ever built. The organization does not decide in any meeting to be less ambitious or to empower fewer people; it simply makes a thousand sensible choices to stay under budget, and the sum of them is a workforce in which a fraction use AI and the rest watch.

The most capable technology of the age is being rationed to a few, not because it fails, but because no one can predict what letting everyone use it would cost.

Copyright © 2026 Concierto. All rights reserved.

6. Predictable Cost Frees Ambition

For the most demanding institutions the argument is not about ambition but about requirement. A systemically important organisation, a national exchange, a central bank, a large regulated financial institution, cannot run on an intelligence layer whose cost is unpredictable, whose outputs cannot be attributed, or whose availability depends on an outside party. For them predictable cost, continuity, and attribution are not preferences to be weighed against price; they are absolute procurement requirements, and sovereignty is binary, the whole estate qualifies or none of it does. What determinism offers the ambitious enterprise as freedom, it offers the regulated institution as the only acceptable way to operate at all.

Make the cost predictable and owned, and both ceilings lift at once: the ceiling on ambition, and the ceiling on how many people are allowed to create.

This is what financial determinism means, a cost the organization sets and can predict, so that ambition and coverage, not the invoice, decide what gets built and by whom. When using more intelligence no longer means paying more, AI reaches the whole workforce rather than a rationed few, and the open-ended work an unpredictable bill deterred becomes possible again. The saving is real, but it is the smaller half of the return. The larger half is the expansion in who gets to create: a fixed, lower cost puts the whole workforce on AI rather than a rationed few, which is affordable at scale only when cost is predictable and owned rather than metered and escalating.

Predictable cost is the precondition for the very breadth that produces the returns, and ownership is what makes the cost predictable.

> **Make the cost predictable and owned, and the question stops being how few people you can afford to give AI. It becomes how much your whole workforce can build.**

7. What Leaders Must Do

- **Price for predictability.** Choose a cost model the organization can forecast, so no team rations its own ambition, or its own people, against an unknown bill.
- **Put the whole workforce on AI.** Pursue the broad, continuous deployment that produces the largest returns, which a predictable, owned cost makes affordable at scale.
- **Invert the order: value first.** Deploy AI on the existing estate for value on day one, and let the productivity fund the technical migration rather than waiting for it.
- **Refuse the token trap.** Recognize that a rented meter looks cheaper early and diverges violently as usage grows, and choose a fixed, owned cost before the divergence arrives.
- **Measure authors, not seats.** Track how many of your people are empowered to create with AI, because that number, not the license count, is what compounds into results.

Copyright © 2026 Concierto. All rights reserved.

A predictable, owned cost is the cheapest way to buy the most valuable thing: a whole workforce free to build, instead of a few permitted to.

8. The Choice

The debate about AI cost is usually about the size of the bill. The deeper question is how many of your people the bill allows to think with AI at all.

An unpredictable, metered cost answers that question with rationing: AI reserved for the few who can justify a bill that rises every year, and a workforce that mostly watches. A predictable, owned cost answers it differently: the whole workforce empowered, the cost held to a controlled, predictable climb as usage grows, and the productivity of thousands of authors funding the transformation around them. The choice is not merely between two prices. It is between an organization that rations its imagination to a few, and one that puts the whole of its people to work on reinventing itself.

Own the cost, make it predictable, and the question stops being who can afford to use AI. It becomes what your whole workforce can imagine.

About This Series

Sovereign Perspectives is a research program on the ownership and governance of intelligence, published by Concierto. Its papers examine the strategic, economic, and institutional questions that artificial intelligence raises for governments and other accountable institutions, and they are written for the leaders making those decisions. The program's aim is to advance the public argument, not to promote a product.

Sources

- Company-reported results, drawn from public earnings statements and reporting, 2024–2026. Figures are company claims rather than audited results; the customer-service example was publicly rebalanced in 2025 on quality grounds.
- The Acme Enterprises figures are an illustrative internal model for an enterprise of roughly 3,000 workloads; specific figures gate to the finance office and are not published as commitments.
- Stanford Institute for Human-Centered AI (HAI), 2025 AI Index Report. Inference cost and adoption context. hai.stanford.edu.

CONCIERTO | SOVEREIGN INTELLIGENCE

reach@concierto.cloud

www.concierto.cloud

www.trianz.com

Copyright © 2026 Concierto. All rights reserved.