

Governing AI: from the cost of the token to the price of the decision

Why AI becomes governable only when a decision binds its cost, its accountability, and its value

The enterprise buyer today hunts the cent per token. They compare rates per million tokens, monitor GPU hours, arbitrate between a frontier model and an open model according to a unit-cost grid. The entire FinOps discourse applied to AI rests on this single operation: counting calls. And counting calls governs one thing perfectly: infrastructure consumption. It governs nothing beyond that. Least of all the decision.

This article extends a reflection on the hidden costs of deployment, where supervision was identified as the institutional cost that the economic models of AI omit. It must be said at the outset, so as not to mislead the reader: what follows does not speak of the cost of AI, but of the way an enterprise transforms a technical event into something it can at last measure, impute, and govern.

Let us set down the definition that holds up everything else. **A cost is governable when an actor holds the institutional lever to observe it, price it, and impose the associated behavior.** The enterprise holds this lever over AI because it controls the point of access to models. But this control is the enabling condition of governance, never its sufficient condition. Holding the door is necessary. It is never enough.

The thesis is broader than a question of cost. An organization governs a phenomenon economically only when it can produce an object where consumption, accountability, and value become simultaneously commensurable. This act of tying, the *binding* (*le nouage*), is the mechanism; governance is its consequence. The decision is today the first operable instantiation of this binding in AI systems, because it is the first object where these three dimensions, hitherto dissociated, are tied together.

The domain of validity is deliberately narrow. The thesis holds neither for conversation, nor for ideation, nor for fully automatic systems, where no human decision is attested. It presupposes an organization endowed with owners, a hierarchy, and accountability. Outside this frame, it ceases to apply. This restriction does not weaken it; it solidifies it.

1. Defining governability, then going beyond it

The token fails as an instrument of governance not because it measures badly, but because it measures the wrong object. It observes a consumption with real precision. It does not price the value produced and imposes no behavior correlated with that value. Of the three verbs in the definition, it serves only one. To observe without pricing or constraining is not to govern; it is to bill. FinOps governs infrastructure. It does not govern the decision.

To understand why the enterprise *can* govern, let us turn to two authors the technical discourse ignores.

- Coase (1937) explains *why* the firm internalizes an activity: because it reduces transaction costs that the market would otherwise make it bear.
- Williamson (1975, 1985) explains *how* it governs what it has internalized, through governance structures adjusted to the attributes of the transaction (specificity, uncertainty, frequency) and calibrated against opportunism.

Coase says why the firm exists. Williamson says how it holds. Our subject lodges between the two. The point of access to models is a governance structure in Williamson's sense: the frontier where the firm can observe, price, constrain. But a structure governs nothing until it has an object on which to act. Exercising the lever presupposes a unit: something one can observe once, price once, impute to someone. Governability is the door. The accounting unit is the room behind the door.

2. From a logic of resource to a logic of commitment

Searle (1995) distinguishes brute facts from institutional facts.

- A **brute fact** exists independently of any institution: a mass, an electrical impulse, a network call.
- An **institutional fact** exists only because a collectivity grants it a status: a signature, a contract, a decision. Its formula, "X counts as Y in context C," applies word for word: a model output counts, or does not count, as a decision.

One consequence of Searle is that the institutional fact exists through collective recognition, not through recording. The decision does not become institutional because a register records it, but because the organization agrees to grant it that status. The register does not invent the decision; it observes a recognition produced elsewhere. To believe the reverse would make the ledger a quasi-magical object, producing the reality it merely records.

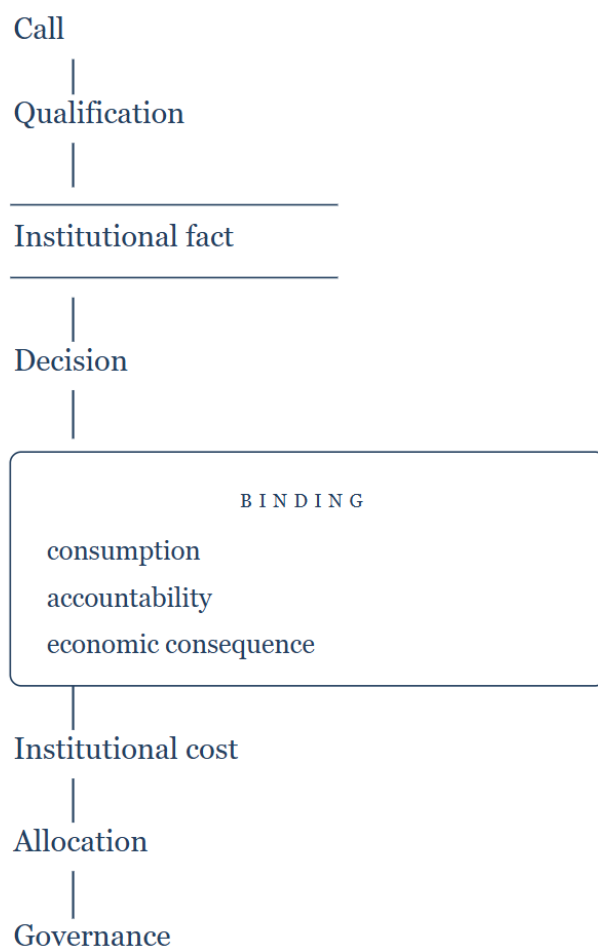
This is the displacement performed by the enterprise that governs its AI.

- A call to the model is a brute fact: it is *consumed*.
- A validated decision is an institutional fact: it is *endorsed*.

The passage is not an accounting refinement; it is a change of logic: one ceases to measure a *resource* and begins to measure a *commitment*, understood as the act that modifies the whole set of organizational obligations attached to a result. The relation runs both ways: deciding what counts as a decision is already, in itself, a governance choice. A call is consumed. A decision is endorsed.

The binding. We call binding (*le nouage*) the institutional operation by which an organization ties, within a single governable object, a consumption of resources, an organizational accountability, and an economic consequence. The binding is the mechanism that renders these three dimensions commensurable. The decision is today its first operable support in enterprise AI systems.

3. The binding, and why it renders commensurable



The core of the theory is not the decision; it is the operation that constitutes it. Why the decision, and not the workflow, the case file, the result? Because it is the only moment where a technical consumption becomes irreversible, converted into an organizational commitment, on one condition that is part of the thesis: that a human endorses it. Before the binding, the three dimensions live in three registers with no common denominator: consumption in the technical logs, commitment in the org charts, value in the operating results. The decision ties them into a dated and attributed object. It is this binding, and it alone, that renders them commensurable. The decision is simply the first object where it appears clearly; the theory bears on the operation, not on the object (Figure 1).

Espeland and Stevens (1998) define commensuration as the comparison of different entities according to a common metric, and remind us that producing it is a labor, and an

exercise of power. An exercise of power calls a question: who, in the organization, holds the right to qualify? The executive committee, the IT department, risk, the business line, legal? This is a question of organizational sovereignty that this article opens without settling. The decision is therefore not a primitive unit (a fragile ontological claim) but a *unit of commensuration*: an accounting unit, because it renders comparable what was not.

From this binding follow the five properties of an accounting unit, as consequences and not as axioms: identifiable, attributable, closable, auditable, additive. Each eliminates a rival unit.

- The **prompt** carries a consumption but no realized value, and stops at the output while economic uncertainty remains entire: it is not closable.
- The **workflow**, diffuse, dilutes attribution: it is not identifiable.
- The **case file** passes from hand to hand: it is not attributable.
- The **result** arrives too late, disconnected from the instant of adjudication: it is not auditable at the right moment.

Knight (1921) grounds closure: ex ante uncertainty is settled only ex post; Hayek (1945) adds that the knowledge to evaluate it is dispersed and reveals itself only in use. One therefore does not price a decision; one settles it.

Additivity, taken literally, would be false: one does not add decisions, heterogeneous in their risks and horizons; one adds their institutional cost, which the binding has reduced to a common metric. This reading falls within a historical regularity whose mechanism is precise: it is not technology that imposes a new accounting unit, it is the exhaustion of the old one, which dies when it ceases to explain the creation of value.

- The labor hour gave way when value departed labor time.
- Activity emerged when value lodged itself in processes (Johnson and Kaplan, 1987).

Today, value lodges itself in adjudication. Simon (1947) had made the decision the unit of analysis of the firm under bounded rationality, where it was an *organizational* resource. What AI adds is that it also makes it an *economic* resource.

The thesis produces predictions, which distinguishes it from a posture:

1. **P1**: at identical compute consumption, two organizations with different levels of institutional qualification present different institutional costs.
2. **P2**: at constant qualification, the institutional cost varies independently of the inference cost; lowering the latter does not mechanically lower the former.

3. **P3:** the gains from technical optimization exhibit diminishing returns as long as institutional qualification remains absent.

4. Three orthogonal layers, one of them central

The ambient discourse conflates into a single block three distinct operations that must be held orthogonal:

1. Route to the right model,
2. Qualify what counts as a decision,
3. Allocate accountability.

Layer 1, technical optimization, is a solved problem. FrugalGPT (Chen, Zaharia, and Zou, 2023) and RouteLLM (Ong et al., 2024) show that task optimization is a commodity. This clarifies an immediate objection: enterprise assistance suites already function without a ledger of decisions, because they operate in the domain of the *assistance request* and leave uncouneted the institutional cost of the decisions they touch. They are not mistaken; they solve a different problem. Layer 1 also fulfills the function Akerlof (1970) names: absent a reliable signal, the demander falls back on model size as a proxy, and the door acts as the screening intermediary that dissipates this asymmetry.

Layer 2, institutional qualification, is the heart of the apparatus, more than layer 3. It answers a binary question: does this event acquire an institutional status, is it a decision? Everything hangs there. Without qualification, no inscription, hence no imputation, no attestation, no accountability. The ledger is a *boundary object* in the sense of Star and Griesemer (1989): a single artifact through which finance, which wants a cost, legal, which wants traceability, and the business, which wants a utility, cooperate without sharing the same representation.

Layer 3 is the allocation of accountability. Orthogonality imposes a guardrail: the ledger records; it does not govern. The sequence is strict: register, then measure, then management decision, then governance. And it is not an ERP; the moment one lets it qualify, trace, and trigger everything, it swells into a platform.

5. The ledger, implementation of layer 2

The ledger implements layer 2, nothing more. A decision is an AI output whose acceptance transfers accountability to an identified owner. This definition draws a watertight boundary. The assistance request (drafting an email, summarizing a meeting) transfers no responsibility: it is out of domain, absorbed by layer 1 without inscription. The act of decision (choosing a treatment, validating a supplier, setting a credit score) commits the enterprise and requires a human owner: it is inscribed, with a balance.

The institutional cost is not the value of the decision; it is what the organization commits in order to produce it and to stand by it. It aggregates two realized costs and a provisioned exposure. Two realized costs:

- **Compute**, the inference of the models mobilized,
- and **supervision**, the human time of validation, graduated according to criticality.

One provisioned exposure: the risk charge the owner accepts in endorsing, which is not an expense already made but a risk that may never be paid.

It is therefore treated as a **provision for decision risk**, estimated on entry, adjusted periodically, settled when one finally knows what the decision actually produced. It is this provisioning that keeps the ledger from being a mere archive: in a long-dated credit or a clinical validation, the real settlement may occur years after the act, but the provisioned line remains steerable in the meantime.

A ledger is not yet a method. Observing compute is trivial, through inference logs; observing supervision presupposes measuring the human time of validation; observing exposure presupposes a provisioning model. A deployed case should document three magnitudes: the **supervision ratio**, expert time relative to criticality and to the model's confidence score; the **friction rate**, the proportion of outputs rejected or deeply reworked, revealing the cost that layer 1 offloads onto the human; and **accounting convergence**, the gap between the exposure estimated ex ante and the cost of the realized risk ex post, over a sample of settled cases. The mechanism is specified; the measured instance remains to be produced. *[CASE TO DOCUMENT: real deployment, even partial, credit granting or clinical validation]*

The choice of word is not neutral. A *register* would be a journal; what is at stake is an *account*, one that aggregates, reconciles, and budgets the institutional cost of decisions. One does not keep the history of decisions; one keeps their account. It is this additivity, applied to cost and not to the decision, that turns traceability into an instrument of management.

6. Allocation of accountability, ex post adjudications, and limits

Layer 3 is the institutional tipping point. Human attestation is not a satisfaction rating granted to the AI, but a choice of endorsement. If the manager validates and endorses, they trigger the pricing of the institutional cost. If the output is imperfect but reworked, the correction time counts as iterative supervision, revealing the real cost that layer 1 transfers onto the human. Only the output neither endorsed nor reworked settles as

technical waste, imputed to infrastructure. The rule that closes the apparatus holds: one cannot use a deliverable and then declare it had no value.

This model shifts the question of usage optimization from a logic of ex ante filtering, founded on a supposed intention, toward a logic of ex post commitment, founded on a commitment actually assumed. Here dissolves a widespread temptation: to govern AI by throttling access according to the supposed laziness of the prompt, through an a priori filter that degrades the model for convenience requests. This path prices intention where it should settle commitment, and it is punitive: it judges the user before their act, whereas the ledger counts it after, with no opinion on them. Throttling laziness manufactures resentment and flight, and revives the very shadow AI it claims to reduce; pricing the decision produces governability.

The same logic settles the case of convenience. If evaluation reveals that a department consumes compute massively without ever inscribing a decision in the ledger, this is not a fault to be sanctioned; it is an accounting fact to be read: a consumption without endorsement, which is requalified ex post as an operational convenience cost and attached to the responsibility center concerned. The apparatus does not judge this usage before the act; it observes it after. The difference from the punitive temptation is not one of degree; it is one of nature.

There remains the most frequent objection, a philosophical one, for it targets the project itself. A good decision, one will say, is a judgment, not an accounting unit. It would hold if one were accounting for the judgment; that is not the case. One does not attach a number to the decision-maker's discernment, but to the commitment their decision creates. Judgment remains irreducible and non-priceable; only its institutional footprint (what the decision consumes, exposes, and produces) becomes commensurable.

A solid theory also says why it fails. The ledger fails if qualification is blurry and lets everything in or everything out, if no decision has a named owner, if the granularity is wrong (counting micro-adjudications or entire case files rather than decisions), or if its implementation cost exceeds the institutional cost it allows to govern. These failures do not refute the theory; they delimit its conditions of exercise.

The apparatus has other limits that must be named. It governs the cost and the accountability of the decision, not its quality: pricing responsibility is not guaranteeing discernment. The door is necessary, not sufficient: quality also depends on workflows, incentives, and rights. And the field excludes conversation, ideation, fully automatic systems, and organizations without a clear owner or accountability (a decentralized autonomous organization piloted by code, an open source collective, a pure cooperative), where the unit loses its attributability. Finally, the theory remains ahead of its empirics.

A word on scope, so as not to confuse it with its proof. This article has demonstrated one case: the decision as the first operable instantiation of the binding in AI systems. But the

criterion is general. An organization governs a phenomenon economically only when it can produce an object where consumption, accountability, and value are tied into a common metric. The decision is today the first such object; others will follow without the criterion changing. The mechanism is broader than its first proof.

Conclusion

All of this holds on one condition, a single one: the enterprise controls the point of access. It is the one place where the door exists by default. Remove the door, and there remains neither routing, nor accounting, nor attestation. This is the situation at the scale of the ecosystem, where no actor holds this position by default, and where governability will have to be created rather than presumed. That is the object of the next article.

Enterprise AI becomes governable neither when one understands it, nor when one counts it by the token. It becomes governable when a call ceases to be a consumption and becomes a commitment for which someone answers.