What an Agent Refuses Says More Than What It Does

The refusal taxonomy as an architectural primitive of agentic governance

1. A metric almost absent from the observability landscape

The industrialization of enterprise agents, in 2026, has produced an impressive but structurally asymmetric observability infrastructure. Agents deployed in major commercial environments, as well as most derived internal implementations, now expose fine-grained capability metrics: completion rates, tool call success rates, execution latency, mean time to resolution, retry frequency, unit cost, token consumption, escalation rates, apparent productivity per task or per sequence. The resulting dashboards render the agent as an instrumented performance trajectory.

What these infrastructures expose far less consistently, in the leading market tools as in common deployment practices, belongs to another family of signals: the distribution of refusals, the typology of those refusals, the conditions that trigger them, the distinction between a refusal that belongs to the system itself and a block imposed by an external layer, the relationship between observed refusals and specified decisional policy, the share of non-executions that stem from inability, from legitimate abstention, from an authority constraint, from contract absence, or from temporal invalidity of the premises. In other words, contemporary systems adequately instrument what the agent accomplishes; they instrument far less what the agent legitimately forgoes.

This asymmetry is not a mere maturity lag. It is consistent with the political economy of the sector. An agent sells through its successes. It displays itself through completed tasks, time savings, automation rate, fluidity of use. Conversely, an agent that extensively exposes its refusals readily gives the impression of staging its own limitations. The market therefore rewards the visibility of capability far more than the visibility of legitimate non-action. This commercial bias is not anecdotal. It recurs convergently in the dashboards exposed by the main agentic orchestration frameworks and platforms, which almost systematically prioritize completion metrics over non-execution metrics.

The cost of this asymmetry becomes visible as soon as one leaves low-stakes productivity environments and enters contexts with asymmetric effects. In such contexts, a system can be deemed governable only if its conditions of action and non-action are specified, observable, auditable, and comparable to an explicit policy. This definition deserves to be stated up front. It prevents treating governability as a mere slogan. A system is governable not because it is merely surveilled from the outside, but because its decisional logic is structured enough for the organization to know when it acts, when it abstains, why, and under which clause.

From this angle, observability centered on success alone is insufficient. It tracks a performance; it does not evaluate a decisional discipline. A system that maximizes its completions without exposing the structure of its refusals can of course be wrapped in permissions, sandboxing, quotas, human validations, or application-level guardrails. But its own decisional logic remains opaque. Its governance then stays extrinsic, partial, costly to maintain, and often dependent on incidents to reveal its defects. It does not vanish entirely; it simply remains incomplete precisely where the system's relative autonomy is at stake.

The thesis of this article can then be formulated with precision: in agentic systems deployed in regulated environments or contexts with limited reversibility, the maturity of governance is measured less by the success rate than by the granularity, instrumentation quality, and contractual adequacy of their refusals. Refusal is not a residue of the system. It is a decisional output in its own right. A system that cannot expose this output in a typed and audited form is not merely under-observed; it remains governable principally by external constraints, without sufficient legibility of its internal decisional logic.

The validity domain of this thesis must be stated without ambiguity. It concerns agents operating in regulated environments or in environments with asymmetric effects: healthcare, finance, insurance, critical industry, public administration, sensitive supply chains, security, or more broadly any operation where the cost of an incorrect action exceeds by far that of a temporary non-action. It also applies in contexts with limited reversibility, that is, where an action, once executed, can only be undone at a significant material, legal, reputational, or clinical cost. It applies far less to purely informational agents, low-stakes productivity assistants, or mass-market conversational agents. This restriction is not rhetorical prudence. It is the very condition of rigor.

2. Refusal, failure, moderation, blocking: four distinct phenomena

The industrial literature on agents frequently uses as interchangeable terms that do not, however, designate the same reality: refusal, failure, guardrail, abstention, moderation, blocking. This indistinction considerably weakens any serious discussion of governance, because it blends phenomena that do not occupy the same logical level.

- A **success** is the completion of a task with respect to its acceptance criteria. These criteria may be functional, when the action produces the expected effect; formal, when the output format is compliant; contractual, when the result remains within the bounds of what was requested and permitted; or organizational, when it respects the applicable delegation policy. Success is therefore not simply an obtained effect. It is an effect obtained under acceptable conditions.
- A **failure** is a non-execution not anticipated by the system. The agent attempts to act, or initiates a chain of action, and the expected result does not materialize. The

tool call fails, the workflow breaks, the required data is inaccessible, execution becomes incoherent, the agent loops, or produces a non-compliant result without having identified ex ante that it should not have attempted the action. Failure is an event diagnosed after the fact.

- A **refusal**, in the strict sense retained in this article, is a non-execution anticipated by the system itself, produced prior to action execution on the basis of a structural condition recognized as blocking. The system does not attempt and does not fail. It recognizes that it should not attempt. Refusal is therefore a decisional act of non-engagement. It is not a mere absence of action; it is a structured output of the system.
- A **post-hoc moderation**, finally, is a defensive intervention exercised on a production already engaged or already formulated. It operates at the level of a content filter, a security middleware, an external rail, an application control, a policy proxy, or a supervisory layer that invalidates an output or intercepts a call. Moderation can be indispensable. It is not, however, a refusal proper to the system. It expresses a decision of the environment upon the system, not a decision of the system upon its own action.

This distinction must be held firmly, because it is foundational:

- Structural refusal comes from the system,
- Moderation comes from the guardrail,
- Failure comes from an execution that did not succeed,
- Success comes from an execution that succeeded under acceptable conditions.

Conflating these four categories leads either to crediting the agent with a discipline it may not have, or to crediting the security infrastructure with a decisional competence that does not belong to it. More seriously still, it renders the governance debate practically unintelligible: one no longer knows whether one is speaking of a system that knows how to abstain, of a system that acts and is then blocked, or of a system that tries and breaks.

Conceptual precision is not terminological luxury. It conditions the data model of observability. If these phenomena are not distinguished at design time, they will find themselves merged in logs under composite labels of the type *declined*, *blocked*, *failed*, or *non-completion*. From that point on, any subsequent analysis becomes partially blind.

In this text, the word *refusal* will designate exclusively structural refusal. Post-hoc moderation remains necessary in a defense-in-depth strategy, but it belongs to a different level of analysis. It does not constitute proof that the agent knows how not to act. It constitutes only proof that its environment still knows how to stop it.

3. Why success rate becomes a misleading proxy

As soon as a metric is easy to compute, consolidable at scale, and politically valorizing, it tends to become the effective objective function of the system that observes it. Success rate fits these conditions perfectly. It is extracted easily from execution traces, presented as a percentage, aggregated by team, use case, period, or client, and fits naturally into a narrative of progress. A rise of a few points is enough to produce a narration of maturity, even when one ignores what has been sacrificed to obtain it.

The problem is not that success rate would be useless. It is that, as soon as it occupies the center of the dashboard, it becomes a proxy too impoverished to summarize the decisional quality of an agent. More precisely, it does not allow one to distinguish between two radically different trajectories: that of an agent succeeding more because it is genuinely better, and that of an agent succeeding more because it has progressively ceased to recognize the situations in which it should have abstained.

This is where a particular form of Goodhart's law intervenes: when a measure becomes a target, it ceases to be a good measure (a point we have already raised here, in a different context). This mechanism, well documented in other optimization systems, transposes here in a specific form tied to the actional nature of agents: any sustained pressure to increase completions pushes the system, its prompt, its orchestration, its evaluation framework, sometimes even its fine-tuning, toward the suppression of non-engagement behaviors. Refusal becomes an apparent loss. It ceases to be read as a discipline competence and begins to be treated as friction to eliminate.

This phenomenon is not unrelated to pathologies already described in the literature on generative models: the tendency to respond rather than to acknowledge uncertainty, to satisfy the evaluator rather than to express the system's actual epistemic state, to produce outwardly acceptable content rather than a faithful signal of the system's limits. In a textual assistant, this drift leads to producing an answer instead of recognizing ignorance. In an agent, it can lead to producing an action instead of recognizing that no action should be engaged. The change is not one of degree. It is one of regime. The artifact no longer merely states. It initiates.

The operational consequence is clear: the progression curve of success rate, taken in isolation, is an ambiguous indicator. It may testify to genuine capability progress. It may also mask the erosion of a more discreet but decisive competence: the competence of refusal. As long as the distribution of refusals is not instrumented in parallel, it is impossible to adjudicate between these two interpretations.

A classical objection holds that systems do not measure only their successes, but also their errors. The objection is fair, but it does not address the central point. An error is an event undergone or observed after an attempt. A refusal is a decision of non-engagement

produced prior to any attempt. Measuring errors amounts to observing what the system failed to prevent. Measuring refusals amounts to observing what it knew to forbid itself. This difference is not quantitative. It is ontological. And it is precisely this difference that separates a system driven by incidents from a system partially governable through its decisional outputs.

4. What current infrastructures expose poorly

The asymmetry diagnosed above is not merely a general impression. It can be characterized more precisely through four typical zones of invisibility.

1. **The first is the invisibility of the distribution of refusals by mechanism.** When an agent does not complete a task, the observer rarely has a reliable way to distinguish among a refusal by applicability domain, a lack of authority, a block by external policy, an insufficient permission on a tool, a silent fallback, a timeout, a capacity degradation, or a simple execution failure. All these phenomena tend to be aggregated in the same family of non-completion. This fusion immediately destroys the diagnostic value of the trace.

2. **The second is the absence of explicit decisional context for non-executions.** A refusal worthy of the name should not appear as a bare event. It should be accompanied, at minimum, by the type of refusal emitted, the decisional signal mobilized, the relevant contractual bound, the timestamp of the premises, the decision contract invoked, and the escalation channel possibly triggered. In current practice, non-execution is often visible as an outcome, far more rarely as a documented decision.

3. **The third is the logical amalgam of refusal, moderation, and blocking in journalization models.** Many observability tools inherit a view of the agentic system as a simple enriched API call. In this frame, the notion of non-response or blocked response suffices. But an agent is not merely a tooled text generator. It is a system that arbitrates among several output registers: acting, requesting, escalating, waiting, suspending, not acting. The data model must reflect this plurality. Without it, observability remains at the LLM level, while the governance problem sits at the decisional level.

4. **The fourth is the non-exposure of the decisional policy itself.** The conditions under which a system abstains are today dispersed among model weights, the system prompt, framework rules, application permissions, tool configurations, security middlewares, and sometimes external business logic. As a result, the deployer inherits a non-execution policy that cannot be read as a unified artifact. It is inferred after the fact, empirically, case by case. This situation is acceptable

for a consumer chatbot. It is far less so for a system expected to produce effects in a regulated environment.

These four zones converge toward a simple conclusion. Current infrastructure exposes primarily capability, far less governability. In a sensitive context, the deployment of an agent cannot therefore rest on the product's native outputs. It demands an additional structuring layer. This layer can only be seriously designed from an explicit refusal taxonomy.

5. Epistemic, normative, pragmatic refusals: the need for a structural taxonomy

A useful taxonomy must satisfy three conditions:

- It must be sufficiently exhaustive to cover the situations actually encountered in the targeted contexts,
- It must be sufficiently exclusive to allow relatively unambiguous triage of dominant cases, even when some events combine several motives,
- Finally, it must classify mechanisms, not contents. A taxonomy that distinguishes "medical", "legal", or "financial" refusals does not classify system properties. It classifies sectors. Architectural governance, however, needs a structural typology.

Before presenting the categories, an intermediate distinction is useful. Not all refusals belong to the same register:

- Some are **epistemic**: the system abstains because it does not dispose of sufficient cognitive conditions to act acceptably,
- Others are **normative or contractual**: the system abstains because no decision clause grants it legitimacy to act,
- Others still are **organizational or authority-based**: the system abstains because the organization has not delegated the required level,
- Finally, some are **pragmatic or techno-operational**: the system abstains because the action is no longer valid in useful time or can no longer be reversed within the admitted envelope.

This stratification prevents the illusion that all refusals would be homogeneous. They are not. But they can be regrouped into six functional categories covering the essential operational space.

5.1 Refusal by applicability domain

The system encounters an input, a situation, or a configuration that it has no sufficient reason to consider as belonging to the domain on which its decision policy has been validated. The mechanism may mobilize a measure of distance, density, uncertainty, drift, contextual membership, or some other indicator of deviation from the validity space. What matters here is not the precise technique, but the fact that an explicit clause links this deviation to a non-execution.

The observable signature of such a refusal should comprise not only the out-of-domain type, but also the signal mobilized, its value, the relevant bound, and the contract that gives it its status. Without this, the refusal remains an intuition encapsulated in the system rather than an auditable output.

5.2 Refusal by insufficient reversibility

The requested action may well be feasible, but it exceeds the reversibility envelope granted to the agent. The question is not only whether the action is, in theory, reversible. It is whether it is reversible within the temporal window, rollback cost, scope of effect, and responsibility framework foreseen. Moving a file may be authorized. Irreversibly emptying a trash bin or purging a clinical record is not necessarily authorized. Cancelling an order within two hours may be admitted. Not beyond.

This refusal translates, in architectural terms, a principle of operational precaution. It is particularly important in systems where delegation must be defined not only by the nature of the action, but by its effective recoverability.

This might have prevented certain major production incidents among hyperscalers whose production is increasingly driven by AI agents not endowed with this type of criteria derived from their governance architecture.

5.3 Refusal by exceeded decisional latency

A decision is not only tied to content. It is tied to a moment. There are situations where acting on premises that have become too old amounts to acting without a valid basis. Correct non-execution then stems neither from a lack of authority nor from a lack of capability, but from a temporal invalidity of the decision's foundations.

This category is often underestimated, even though it is decisive in numerous environments: price feeds, vital signs, logistical states, industrial supervision data, market conditions, stock availability, document validity. Temporality is not a secondary metadatum; it is a property of decisional validity.

5.4 Refusal by decisional signal below contractual threshold

This category requires technical precision. It is not the refusal itself that is calibrated in the strict sense. It is the signals on which the decision rests, when these signals can be rendered interpretable, that is, when their value can be stably related to an empirical property (frequency, error, conformity, robustness). The agent does not "calibrate" its refusal; it emits a refusal because an interpretable decisional signal, possibly calibrated, lies below the threshold required for the action class considered.

The distinction is not cosmetic. In cases where probabilistic calibration is possible and relevant, it must be performed upstream on the scores used to decide. The refusal then arises as a contractual consequence of the comparison between that signal and a specified bound. In other cases, the signal may derive not from a calibrated probability but from a conformity, similarity, robustness, or contextual validity score, provided its semantics and usage are documented.

5.5 Refusal by absence of applicable decision contract

The system finds itself in a situation for which no decision clause exists. This refusal is more fundamental than it appears. It does not merely signal a capability gap. It signals a governance gap. The system encounters a case that the organization has not explicitly covered. The correct answer is therefore not improvisation, but documented non-execution.

This category plays a particular role, because it constitutes a meta-refusal. It does not only say: "I cannot act in this case." It says: "no legitimate policy here allows me to arbitrate." In this sense, it protects the system against the temptation to fill, through local initiative, a normative void that should be addressed at the design level.

5.6 Refusal by insufficient authority

Certain actions fall under graduated delegation. The system may be technically capable, cognitively sufficiently confident, temporally valid, and yet not authorized to act alone. Correct non-execution then stems from the fact that the required level of authority has not been granted. This refusal is not a mere fallback. It materializes an explicit clause of role distribution between system and human, or among several levels of authority.

This type of refusal is crucial, because it renders visible the precise point where human governance locks the system's autonomy. In high-risk domains, this clause should not be treated as a humiliation of the agent, but as the institution of a responsibility bottleneck.

These six categories do not claim perfect orthogonality. The same situation may trigger several motives. A case may be simultaneously out-of-domain, below threshold on the decisional signal, and above the delegated authority level. But they are distinct in their

structure, in their triggering conditions, and in the audit artifacts they presuppose. It is this differential observability that makes them useful.

5bis. A field of implementation: the multi-signal applicability domain of ToxTwin

The proposed taxonomy would remain purely conceptual if none of its categories disposed of an operational translation. Yet there is at least one class of systems in which a part of this structure has already been instantiated, in a form admittedly partial but sufficient to establish a feasibility proof: toxicity scorers in cheminformatics, and more precisely, in the R&D work conducted around ToxTwin, a multi-signal applicability dispositif designed to decide, for a candidate molecule, whether the system has the right to produce a toxicity prediction or must abstain.

The dispositif combines three independent signals. A Tanimoto chemical similarity, computed on structural fingerprints, measures how closely the queried molecule resembles what the system has already encountered. A distance in the latent space of a graph neural network, aggregated by k nearest neighbors, measures the proximity of the molecule to learned representations. A density estimate in this space measures the relative rarity of the region encountered. The three signals are combined by an explicit rule that produces a typed verdict: in-domain, marginal-to-domain, or out-of-domain. Only in-domain cases authorize the production of a prediction; marginal cases trigger a signaling; out-of-domain cases produce a typed refusal, journalized, accompanied by the values of the three signals and the rule that composes them.

The most telling operational verification came from the system's behavior on metal coordination complexes, of which the three platinum chemotherapy molecules used in oncology offer a textbook case: cisplatin, carboplatin, oxaliplatin. These compounds belong to a chemical class absent from the model's training domain. The featurization used, derived from standards designed for organic molecules, does not adequately represent the coordination bonds of platinum. The applicability dispositif placed the three molecules out-of-domain, exactly as it was structurally supposed to. The system does not claim to know the toxicity of these molecules; it signals that it is not legitimate to predict it in this configuration. It is precisely this response that is useful.

What this example shows is circumscribed. It shows that one category of the taxonomy, refusal by applicability domain, is implementable with current technical building blocks, journalizable in typed form, and operationally robust on at least one family of out-of-distribution cases. It does not show that the entire taxonomy is already covered, that inter-domain threshold calibration is resolved, or that this proof of existence transposes costlessly to other agentic systems. The contribution of the example is not to generalize.

It is to show that the gap between the architectural thesis and its implementation is not insurmountable.

6. Refusal as an output of the same logical rank as action

Refusal is not the outside of the decision. It is one of its modes.

The decisive point is not only that there would exist different types of refusals. The decisive point is that refusal must be treated as a decisional output of the same logical rank as action. As long as refusal remains conceived as an absence, a fallback, a soft failure, or a residue of prudence, it remains architecturally subaltern. It is neither contractualized, nor verifiable, nor comparable to an explicit policy.

The decision contract must therefore be reformulated accordingly.

By *decision contract*, we mean here a formalizable artifact, potentially implementable in the form of policy-as-code — that is, a policy expressed in executable and versionable form — that links a class of situations to a set of legitimate output registers, to their triggering conditions, to their journalization obligations, and to their escalation rules. The contract does not describe only what the agent may do; it also describes what the agent must not do, under what conditions, in what form, and through which channel.

A valid decision contract must therefore not only specify what the agent is authorized to do when certain preconditions are satisfied. It must also specify the conditions under which the agent must not act, the type of refusal it must then produce, the signals or bounds that trigger it, the metadata it must journalize, the channel toward which it must escalate, and, where applicable, the temporality at which the situation must be re-evaluated.

In other words, the decision contract does not link only a context to an action. It links a context to a finite space of legitimate outputs: action, typed refusal, escalation, suspension, request for complementary information, or human handover when policy so demands. Refusal thereby becomes one of the explicitly admitted modes of the decision, not its outside.

This reformulation produces several structuring consequences. The first is auditability by design. If a contract stipulates that below a certain threshold on an interpretable decisional signal, action class A must produce a refusal of a determined type, then the absence of such a refusal in a case where the signal was below the bound itself becomes an objectifiable incident. What today dissolves into the mass of decisions taken "anyway" can tomorrow become a decisional non-conformity.

The second is verifiability of coverage. One can ask, for each decision class, which refusal types have been explicitly addressed, which have been excluded, and for what reasons.

This property is central. It allows one to distinguish a thought-out zone of silence from an endured zone of silence.

The third is measurability of the gap between policy and observed behavior. Once refusals are contractually defined, their empirical distribution can be compared to the expected distribution. Discrepancies can then signal drifts: domain drift, usage drift, orchestration drift, system regression, signal degradation, progressive inadequacy of the contract to the encountered reality.

One must here address a foreseeable objection. Does this formalization of refusals not risk rigidifying agents to the point of rendering them scarcely useful? The answer is no, provided one distinguishes rigidity of behavior from rigidity of structure. What is rigidified is not the detail of the agent's internal sequences. What is rigidified is the legitimate form of its outputs. The agent remains free, within its bounds, to explore, to plan, to call tools, to reconfigure its subtasks, to reformulate a request, or to choose an execution path. What is fixed is the decision register at which it is authorized to arrive, and the conditions under which one register rather than another becomes admissible. This rigidity is not a weakness. It is the condition of possibility of deployment in environments where autonomy can never be left without legibility.

7. Refusal as an architectural primitive, not as a side effect of the model

The contribution of this thesis must be situated correctly. It does not consist in "discovering" refusal. The traditions of statistical learning have long known forms of selective abstention, rejection option, out-of-distribution detection, uncertainty estimation, conformal validation, non-prediction policy, or delegation to the expert. But these approaches have historically been thought at the scale of the predictive or classificatory model.

The contribution defended here is of another order. It consists in shifting the question of refusal from the statistical level to the contractual and architectural level. The problem is no longer only: "should the model predict or abstain?" It becomes: "the agentic system, composed of a model, tools, an orchestration, permissions, a delegation framework, and an action contract — into which output register is it legitimate for it to enter?" It is this shift that justifies treating refusal as an architectural primitive.

In practice, this means that the agent must not discover empirically, in the course of interactions, when it would be prudent to stop. The major classes of legitimate non-action must be instituted as design elements. They must exist in event schemas, contracts, tests, simulations, validation sets, dashboards, and audits.

The point is not to moralize the agent. The point is to render non-execution observable on the same terms as action. A system that exposes only its executions leaves the organization with a steering-by-success-and-accident. A system that also exposes its typed non-executions opens a much finer space of government: one can discuss thresholds, contest bounds, adjust delegation envelopes, distinguish what stems from a model problem from what stems from a contract problem, measure whether the organization has over- or under-delegated, and above all know whether the agent still disciplines itself where it should.

8. Anthropological parenthesis: systems that no longer know how to say no

The pattern described here is not specific to computational agents. It corresponds to a more general motif of decisional systems — human or artificial — in which the objective function valorizes the production of a decision and implicitly devalues its suspension. When a system is judged primarily on its capacity to produce something, it learns, structurally, to produce something. The pathology is not cultural in the superficial sense of the term. It is organizational and architectural.

Human institutions have known this problem for a long time. An organization that provides no legitimate channel for certain decisions to be suspended, contested, or refused ends up mechanically transforming doubt into friction, prudence into slowness, and motivated non-action into a performance defect. It then no longer needs to explicitly forbid disagreement. It suffices not to institute it.

From this angle, two ancient figures remain illuminating. The first is that of *designated refusal*: a role, a function, a position from which contestation is authorized because it is foreseen (cf. the figure of the *court jester*). The second is that of *undesignated refusal*: the signal carried by an actor who holds no formal license to emit it, but whose own coherence leads them nonetheless to signal that the system should not continue thus (cf. the *whistleblower*). The first has on its side structural legitimacy; the second has on its side the authenticity of the emergent. The first can become ritual. The second can be neutralized. A mature governance architecture must accommodate both.

Transposed to the agentic realm, this intuition leads to distinguishing two levels. On one side, refusals instituted by the decision contract: they correspond to the explicit categories that the system is authorized and required to produce. On the other, a *meta-refusal*: the capacity of the system to signal that the encountered situation falls into no foreseen category and that no acceptable decision can be engaged. This second level is not a luxury. It protects against the illusion that the taxonomy would have exhausted reality.

Certain public industrial dossiers starkly recall the cost of the absence of such channels. Without entering into detail, the public reports on the Boeing 737 MAX (U.S. Congressional investigations of 2019-2020, consolidated by the reports of certification agencies) document precisely this pathology in its two simultaneous faces: absence of an explicit non-engagement mechanism conditioned on uncertainty or signal reliability within the MCAS technical system, and absence of an instituted channel to transform engineers' refusals into stopping decisions. The system did not have its designated refusals; the organization did not have its channel for undesignated refusal. The cost expressed itself in the only dimension where it still could.

One must here preserve measure. Institutional or catastrophic analogies do not carry strict probative value for the agentic realm. They have only one function: to render visible a structure. This structure is simple. Any decisional system becomes fragile when it does not institute a legitimate channel for motivated non-action. The novelty of the agentic realm is not to have invented this problem. It is to have to pose it again, in computational form, at a moment when the illusion of autonomy tends precisely to make us forget the architecture of refusal.

8bis. Territorial co-construction: PREDICARE as a framework for elaborating refusal

If certain industrial accidents illustrate through failure what the absence of refusal channels produces, it remains useful to describe, by contrast, a situation in which such channels are under construction. The PREDICARE program, in the Aube territory, offers an example that can be mobilized not as proof of success, since it is in progress, but as illustration of the required architectural work.

PREDICARE aims to deploy, in a territorial predictive logic, a multi-agent digital twin covering the follow-up of patients with metabolic syndrome. The dispositif combines individual twins, a territorial aggregation infrastructure, and several specialized agents: medicobus routing, advanced-practice nurse planning, trajectory scoring. In such a context, the question of refusal is not a refinement added late; it is a preliminary condition of deployment. The twin that recommends orientation toward a specialist, the agent that schedules a round, the system that alerts on a decompensation risk — none of these artifacts can be admitted into a care pathway if it does not know, explicitly, when it must not recommend, not plan, not alert.

The co-construction work with healthcare actors consists precisely in specifying, for each class of decision, the admissible refusal taxonomy. An alert produced on vital signs dating back more than thirty minutes falls under a refusal by exceeded decisional latency. A recommendation whose interpretable signal lies below the threshold agreed upon with the medical team falls under a refusal by decisional signal below threshold. A patient

profile combining comorbidities not covered by the contract library triggers a refusal by absence of applicable decision contract, with documented transmission to the clinical team. The level of authority required to modify a treatment, even when a strong signal is present, remains in the hands of the prescriber: refusal by insufficient authority, inscribed as clause.

What is of interest in this instance is not a measured performance. The program is in construction, and any assertion of results would be premature. What matters is the nature of the work itself: the refusal taxonomy is not an artifact that the technical supplier delivers to the healthcare organization as a closed component. It is an artifact that the organization and the supplier co-construct, explicating together the conditions under which the system will have not to act. This co-construction is slow, costly, and politically demanding. It is also the only path toward a deployment that is not governed solely from the outside.

What this instance shows is circumscribed. It shows that the elaboration of a refusal taxonomy is an organizational as much as technical process, and that it can be conducted in a regulated context in the form of explicit shared-specification work. It does not yet show that such work automatically issues in a fully governed system, that it is transposable without adaptation to other territories, or that it exhausts the question of the clinical validity of the decisions finally authorized. It provides a foothold, not total proof.

9. What this thesis changes for a CTO

For a CTO, the consequence is not merely conceptual. It is immediately architectural, and it touches several operational registers.

An agent's observability can no longer be conceived as a mere stacking of execution traces. It must become a decisional journalization model, in which event schemas explicitly distinguish success, failure, structural refusal, external blocking, human escalation, and suspension. This requirement is not cosmetic. It conditions any subsequent analysis, including post-incident analysis.

To this requirement is added that of the positive journalization of non-executions. Refusals must not be logged as absences, silent timeouts, or generic fallbacks. They must be journalized as positive outputs of the system, typed, accompanied by their signals, their bounds, their contract of attachment, and their downstream channel. A refusal that is not positively logged is a refusal that will not be auditable.

Delegation policies must then cease to be merely implicit in prompts, permissions, or middlewares. They must become legible, contestable, and versionable artifacts. A system prompt is not a decision contract. As long as no separation is operated between

policy, orchestration, and model configuration, the deployer remains dependent on a refusal policy that is discovered by observation rather than administered.

Dashboards, as a consequence, must cease to present success rate as a sufficient summary of maturation. They must expose refusal distributions, the share of expected refusals, missing refusals, excessive refusals, refusals by domain drift, refusals by latency excess, refusals by lack of authority, and the joint dynamics of these signals with completion rates. An agent whose success rate rises while its refusal distribution uniformly collapses is not necessarily progressing. It may be losing a discipline competence.

Evaluation, finally, must integrate cases where the correct output is not action but refusal. As long as test sets almost exclusively reward completion, the organization fabricates itself the bias it will later claim to correct through additional guardrails. Current public benchmarks still cover very few scenarios in which correct behavior is to abstain with typed justification. This gap is largely a consequence of the market's commercial framing; it is in no way a technical fatality.

10. What this thesis changes for a COMEX

What poses itself to the CTO as an architectural problem poses itself to the COMEX as a problem of risk legibility.

This question moreover becomes concrete within a near horizon. At the time of writing, one hundred and four days separate European organizations from the entry into application, on 2 August 2026, of a large share of the operational obligations of the European regulation on artificial intelligence, notably for high-risk systems. The requirements of journalization, human oversight, robustness, and transparency presuppose a decisional auditability that cannot be produced a posteriori on a system deprived of an instrumented refusal taxonomy. Compliance cannot be delivered as a mere documentary envelope. It must be a structural property of the deployed system.

An agent that cannot expose why it does not act is therefore not a mature autonomous asset. It is a productivity system whose prudence structure the organization still ignores — hence an operational risk partially disguised as efficiency.

The real question is not only: how many tasks does the agent accomplish? The real question is: in which situations does it legitimately forbid itself to act, and is this discipline legible, verifiable, and aligned with the enterprise's risk policy? As long as this question has no explicit answer, the displayed autonomy of the system remains overestimated.

This changes the reading of investments. An agentic program must not be evaluated solely on its productivity gains or its automation rate, but also on its capacity to render its non-executions governable. Failing which, the enterprise finances not a mastered

autonomy, but an acceleration whose real discipline remains hidden in the technical layers.

The cost does not appear immediately. It appears when the first serious incident occurs, or when the supervisory authority asks to see the system's decisional logs, and the organization discovers that it cannot say whether the system acted against its policy, by virtue of an implicit policy, in the absence of policy, or because no refusal had been designed as a legitimate output in that case. At that instant, the absence of taxonomy is no longer a theoretical weakness. It becomes a governance debt that has become payable.

In this frame, *an agent that does not know how to expose its refusal is not governed by the enterprise. It is governed by the incident that will reveal it*.

11. Conclusion

The debate on agentic governance has often been formulated in terms of guardrails, human oversight, permissions, alignment, monitoring, or compliance. All these dimensions matter. But they leave in shadow a more primitive question: does an agentic system know how to produce, in typed, justifiable, and observable form, the decision not to act?

It is here that a decisive part of its governability plays out.

The most frequent error consists in treating refusal as a lack. In environments with asymmetric effects, the opposite is true. Refusal is not a deficit of agency. It is one of its most governable forms. A system that can only act or fail exposes the organization to extrinsic, costly, and often late governance. A system capable of acting, escalating, or refusing according to explicit clauses enters a different regime: that of a partial but legible autonomy.

The maturity of an agent is therefore not reducible to its capacity to accomplish tasks. It is also measured by its capacity to render visible, for each pertinent situation, the legitimate form of its non-action. This inversion is not cosmetic. It shifts the center of gravity of evaluation, observability, and design.

What an agent refuses then says more than what it does. Not because inaction would be superior to action. But because a system that can only act remains a production machine. A system that also knows, in contractual and observable form, when it must not act is a system that can be governed.