Annex I Trilogue: The Track Signs the Dossier, Not the System

The false debate

The European debate on medical devices incorporating AI is currently conflating two distinct objects: the framework that certifies the dossier and the architecture that governs the decision. This confusion structures today's trilogue.

The Annex I trilogue of 28 April 2026 has bogged down on a question that has become central: which regulatory architecture should be retained for medical devices incorporating artificial intelligence.

The debate sets two familiar positions against each other.

The European Parliament, several Member States and part of the sectoral federations defend a single sectoral track. Their argument is coherent: the MDR/IVDR pair, complemented by IEC 62304, ISO 14971, ISO 13485, clinical validation protocols and post-market surveillance, already constitutes a dense apparatus for managing software risk across the lifecycle. Adding a further horizontal layer would introduce documentary redundancy, regulatory friction and industrial slowdown.

The Commission, conversely, maintains the logic of a crossed track MDR + AI Act, on the grounds that certain risks specific to AI systems are not adequately covered by existing sectoral frameworks.

Both positions are technically defensible.

They nonetheless share a single premise, rarely made explicit: that a regulatory track can on its own produce the governability of an AI system in production.

It is this premise that does not hold.

The choice of track determines the dossier to be signed. It does not determine whether the system remains governable at the exact moment it decides.

A runtime property, not a lifecycle property

The problem is not documentary. It is architectural.

A clinical AI system does not engage its responsibility when it is developed, documented or certified. It engages it when it produces an enforceable decision on a real patient, under open distribution, in operational conditions that are imperfectly controlled. It is at that precise moment that governability is at stake.

The governance of a probabilistic system is therefore not a lifecycle property. It is a runtime property at inference.

This distinction is not academic. A system can be simultaneously validated, certified, traceable, monitored and compliant with the applicable framework, and remain incapable of determining, at the exact moment it produces a decision, whether that decision should be *accepted, refused, downgraded, transferred to another computation path*, or *escalated to a human*. Yet it is precisely this capacity that decides whether a system is governable.

The current debate organizes primarily the compliance of the lifecycle. It does not specify the minimal runtime properties that make an AI decision legally and operationally governable.

The blind spot common to both tracks

The two regulatory architectures currently under debate regulate primarily processes, documentary obligations, risk management mechanisms and ex-post surveillance procedures. Neither explicitly requires execution properties auditable at inference.

This nuance is decisive.

A derivable property is not an enforceable property.

A text may allow one to infer that a manufacturer should monitor model drift, control certain inputs or restrict certain uses. As long as these properties are neither explicitly specified, nor auditable in production, nor verified at runtime, they remain interpretable, heterogeneous, hard to audit and legally fragile. The system can then remain compliant while taking ungovernable decisions.

The actual failure mechanism

The failure mechanism observed in probabilistic systems is relatively simple.

An input arrives outside the distribution on which the model was calibrated: a new population, a degraded sensor, a different clinical protocol, a rare biological combination, incomplete data, or silent temporal drift. The system nonetheless produces an inference. The calculated probability is interpreted as an actionable decision. No signal qualification intervenes; no admission filter blocks execution; no fallback route activates. A clinical, prudential or economic decision is taken on a signal whose actual reliability is not qualified.

The failure does not come solely from an algorithmic error. It comes from the absence of architectural properties capable of interrupting the failure chain before an enforceable decision is produced.

The system does not need to know that it is wrong. It must know that it could be.

It is this capacity, distinct from correctness, that separates a governable software asset from a probabilistic operational debt.

The empirical signals

The available data do not contradict this diagnosis.

A study published in 2025 in *JMIR Medical Informatics* (Chen, Teng, Kuo *et al.*, 2025[1]), covering twenty-seven years of AI/ML recalls in the United States drawn from the openFDA base, documents a recall rate of approximately 5.8% on FDA-cleared AI devices, with a predominance of causes linked to software design and device design, these accounting for roughly 50% of the root causes identified in the AI/ML cohort. A complementary analysis on the FDA AI-Enabled Medical Devices List[2] further confirms that only a minority of listed devices publicly document detailed premarket safety evaluations.

These data do not establish a single causal link. They remain nonetheless compatible with a more structural finding: current frameworks know broadly how to audit the software lifecycle; they still insufficiently specify the runtime properties that govern probabilistic decision in production.

The minimal architectural criterion

What does a governance effectively inscribed in the architecture of the system look like?

It looks like a minimal specification of three controls executable at inference: *validity control, admission control, transfer control*. These three functions are not engineering preferences. They correspond to the three points where the failure chain can be interrupted before an enforceable decision is produced. Removing any one of them is enough to reintroduce systemic risk.

Validity control

A governable system must explicitly determine whether the incoming signal still belongs to the zone for which its performance has been demonstrated. This requires validation without methodological leakage, a split respecting the real conditions under which future data appear, calibration verified on an independent set, and the explicit declaration of an applicability domain. Outside this domain, inference must be considered unreliable by default.

The corresponding technical device, the explicit declaration of the zone within which performance remains claimed, constitutes the *validity port* of the system. The grammar of ports was developed in Article V of the hexagonal series; the validity port is its most immediately operational instance for deployments in regulated environments.

The minimal admissible proof is not an additional performance curve. It consists of a published validation protocol, a documented applicability domain, and the expected degradation conditions outside that domain.

A system incapable of qualifying its own perimeter of validity cannot enforce its decision.

Admission control

A governable system must decide whether a request is admissible before the model is even called. The admission decision can be neither implicit, nor statistical, nor left to the end user's interpretation: it must explicitly produce one of three outputs:

1. *acceptance,*
2. *rejection,*
3. *rerouting*

on the basis of publishable and instrumented rules.

The critical point is not the model's ability to respond. The critical point is its ability to refuse.

The minimal admissible proof consists of the published admission rules, the typology of rejections, and the observed rate of refusal and rerouting.

Without admission control, the system does not govern its action space. It undergoes it.

Transfer control

A governable system cannot rest on a single computation route. It must contain at minimum a nominal route, a fallback route to a calibrated reference model, and an escalation route to a human or specialized system. The switching criteria between these routes must be explicitly defined, logged and activatable.

Escalation cannot be an abstract regulatory formula of the "human validation required" type. It must correspond to an execution mechanism effectively activatable. Unrouted human supervision is not a guarantee. It is a compliance narrative.

The Predetermined Change Control Plan (PCCP) imposed by the FDA for SaMDs probably constitutes the first explicit regulatory shift toward a pre-specified governance of the system's evolving behaviors: the model's evolution route is there pre-specified, logged and activatable according to criteria published upstream of deployment. The PCCP, however, only formalizes the governance of model modifications; it does not, on its own, cover the mechanism of decisional escalation at inference. It nonetheless marks its direction. The question is no longer whether a human is watching. It is whether the machine knows when to disturb them.

The minimal admissible proof consists of the published switching criteria, the observed escalation rate, and the response times of the fallback routes.

Without transfer control, a probabilistic system possesses no structural mechanism for containing its own failure conditions.

An industrializable doctrine

This architecture is not theoretical. It already exists in certain industrial multi-model systems where the nominal route is conditioned on prior signal qualification, where a calibrated fallback takes over in cases of uncertainty, and where human escalation is triggered outside the applicability domain.

On PREDICARE, a territorial predictive medicine program structured for the GHT Aube, the reference architecture implements this separation by construction. Every incoming signal, whether from a connected sensor, a hospital information system or an ambulatory actor, is qualified according to a *Bronze / Silver / Gold* typology before any inference. A Bronze signal (a blood pressure reading at 14:37 without verified clinical context, for example) does not cross the admission layer. A Silver signal authorizes an inference on the nominal route. A signal degraded during execution switches to a fallback route on a calibrated reference model. A signal outside the signed applicability domain triggers an escalation to the clinical practitioner of the reference CPTS, within a defined, traceable and auditable time frame.

This instance does not demonstrate that the doctrine is universally satisfied. It demonstrates that a system explicitly carrying the three controls at inference is technically industrializable, and that its reference architecture can be specified in a real regulated environment, and not only in a laboratory demonstrator.

A symmetrical gap: EU / US / CN

International comparison does not modify the diagnosis.

The European framework, whether sectoral or crossed, does not explicitly require the three execution controls. The American QMSR, in force since 2 February 2026, names model drift and data governance without formally requiring the delimitation of the applicability domain, the online filtering of inputs, or multi-path routing at inference. On the Chinese side, the NMPA simplifies change registration procedures as long as the core algorithm is considered stable; there again, control bears primarily on the identity of the model and its modification cycle, not on the execution properties of probabilistic decision.

The gap is therefore symmetrical. It does not reflect a European lag. It reflects the current state of worldwide regulation of real-time probabilistic systems.

Acknowledged cost and friction zones

The three controls are not free.

Validity control imposes a more demanding validation protocol and the continuous maintenance of an explicit applicability domain, a declaration that must be revised at each evolution of the target population, the clinical protocol or the upstream sensor. Admission control adds latency to each inference, which becomes critical on high-volume pipelines or in emergency contexts. Transfer control requires continuous logging and an escalation infrastructure, human or algorithmic, that must itself be available, calibrated and auditable.

Three friction zones deserve to be named without being minimized.

1. First friction: declaring an applicability domain remains technically delicate for foundational generalist models, whose performance zone is by construction blurred and evolving. The doctrine does not eliminate this difficulty; it forces it to be made explicit rather than masked. The benefit is epistemic, not computational.
2. Second friction: admission control can come into tension with the demands of clinical reactivity in emergency situations. The doctrine does not impose a blocking admission by default; it imposes an explicit typed decision, which may include a priority degraded mode that is activatable and logged. Governability is not synonymous with slowness; it is synonymous with traceability of arbitration.
3. Third friction: the total operational cost is measurable, in computation hours, lines of code, and full-time equivalents. The alternative cost, first contentious litigation, first adversarial audit, first critical incident on a mis-qualified signal, is neither measurable a priori nor boundable a posteriori. The choice is not between governability and performance. It is between known architectural cost and unknown incident cost.

The real stake of the trilogue

The trilogue remains useful. It organizes responsibilities, defines perimeters, clarifies obligations and structures European regulatory industrialization. The doctrine defended here does not intend to substitute itself for it; it proposes to extend its scope of specification from the lifecycle toward runtime properties. But a framework cannot resolve a property it has not explicitly specified.

The thesis defended here remains falsifiable. It would suffice for a device compliant with an existing framework to publish an explicit specification of its execution properties, auditable runtime metrics, verifiable mechanisms of qualification, rejection and escalation, and an independent audit under real conditions, to demonstrate that a

current framework already suffices to produce this governability. In the absence of such public dossiers, the conclusion holds.

The track signs the dossier.

Governability plays out at inference.

Without validity control, without admission control, without transfer control, none of the two regulatory architectures currently under debate guarantees that a system remains governable at the exact moment it decides.

This note formally addresses the governance of AI-enabled medical devices in the context of the European trilogue. It poses in the background a broader question, which will be the subject of 4/6: under what conditions does a decision remain enforceable when the engine that produces it is probabilistic and partially non-deterministic? It is to this question, and not to that of compliance, that ***operational governance*** will have to respond.

[Series: Governance = architecture - 3/6].

See also:

Article V (hexagonal architecture, ports and adapters),

The Contractualised Promotion Port: What the FDA Built Before the Architects Named It · Twingital Institute

The RAISE Framework · Twingital Institute