The problem is not imperfection

The debate on the evaluation of clinical artificial intelligences has settled on a finding no one contests: benchmarks are imperfect. That is true, and it is beside the point. No serious practitioner has ever believed a benchmark to be exhaustive, and one does not write a doctrine on a self-evident fact. The problem is harder, and it can be stated in a single sentence, which everything else in this article does nothing but hold, test, and operationalize:

A clinical benchmark defines which forms of failure become governable objects for the system.

That is where the entire stakes are concentrated. The benchmark, understood here in a deliberately broad sense as a validation space, is not only a measurement instrument: it is the perimeter effectively used to decide that a system is sufficiently valid to be promoted, that a behavior deserves to be monitored, that an incident deserves to be treated as a systemic signal. What does not enter this perimeter does not disappear from clinical reality: it exits institutional attention. Perimeter validity is not governance. It is its condition of possibility.

Three successive shifts in the burden of proof

1. The decade 2015 to 2025 saw a first shift. Systems were no longer judged on methodological plausibility alone; reproducible performance metrics were required: AUROC, sensitivity, calibration, F1. That shift was sound but limited, and criticism seized on it rapidly. An average score on a narrow benchmark guarantees neither transferability nor utility.
2. The editorial sequence of 2026 operated the second shift. *Nature Medicine* no longer asks whether models hallucinate, but whether they improve clinical outcomes, and answers, across several convergent texts, that in many cases we do not know (*Is AI actually improving healthcare?*; *Show us the evidence for the value of medical AI*, 2026). The burden of proof passed from intrinsic performance to proof of impact. That is correct, and it is necessary.

But this shift, in its turn, does not suffice. The trap is subtle: even the proof of clinical impact silently inherits the visibility perimeter defined upstream. A randomized pragmatic trial demonstrating an improvement in length of stay or readmission rate says nothing about the classes of incidents that remain invisible to the measurement instrument used in the trial itself. The outcome validates what it observes. It remains silent on what it has not instrumented.

3. The third shift, the one this article seeks to formulate, does not replace the first two. It completes them. It reads as follows: the relevant question is no longer only *is the model performant* or *does it improve outcomes*, but *will the incidents it produces in production remain observable, attributable, and contestable*?

Performant is not governable

Three validity levels are conflated under the word "validation," and separating them is the central contribution of this text.

Statistical validity answers: does the model predict correctly within the benchmark? *Clinical validity* answers: does the benchmark represent the relevant clinical situations? *Governable validity*, the third level, answers a question the first two never ask: will the incidents produced in production remain institutionally manageable?

This validity stratifies into three planes that must be held separate, because they are frequently conflated, and the conflation is costly. *Observable*: the incident produces a trace the system records. *Attributable*: the trace can be linked to a class of incidents, to a model, to an identifiable decision. *Governable*: a stable and contestable organizational policy exists to handle this class.

An incident can be observable without being attributable: something happened, it cannot be qualified. It can be attributable without being governable: the class is known, no procedure triggers. A model's AUROC score says nothing about any of the three planes.

A model can therefore excel at the first level, pass the second, and fail the third without anything in its score signaling it.

An example makes the absurdity of the leaderboard reflex physically palpable. Consider model A, AUROC 0.96 on a narrow benchmark. Consider model B, AUROC 0.91, but with explicit coverage of critical rare classes and a dedicated escalation policy for each. The aggregate performance ranking prefers A. Governability prefers B, because B knows how to recognize the rare incident as an incident, attribute it, trace it, escalate it, where A dissolves it in a flattering average. This is not a paradox; it is what happens when average accuracy ceases to be conflated with contestable safety. The leaderboard reflex optimizes exactly the wrong quantity.

The preference for B must not become a naive apology for over-signaling. A system too eager to recognize classes produces the symmetric pathology: alert fatigue, which destroys governance through saturation. Authentic governability is not the maximization of detectability; it is its calibration. The criterion is not how many events escalated, but which classes of events, under which arbitration policy. Conflating the two replaces a blind spot with fog.

The propagation theorem, and its cost

At this point, the dominant industrial objection presents itself, and it is solid: post-market will compensate. Real-world surveillance, incident tracking, modification management plans will correct downstream what the benchmark missed upstream. The objection deserves to be taken seriously, because it describes a real mechanism.

The response is what can be called a propagation theorem. It must be formulated precisely, because its rough version is attackable and its exact version is strong.

Let S be a deployed surveillance system and T its initial visibility taxonomy. An event class c absent from T is not rendered impossible for S to discover: a weak signal, a clinician report, a pharmacovigilance cluster, a retrospective qualitative analysis can identify it.

What the theorem asserts is more precise and more exact: the absence of c from the initial perimeter conditions the institutional cost of its subsequent recognition as a governable class.

This formulation shifts exactly the point a pharmacovigilance expert would rightly object to. Surveillance systems serve precisely to discover unknown classes. True. But that discovery is not free. A class not instrumented from the outset requires, in order to become governable, a complete institutional reconstruction: new taxonomy, new threshold, new protocol, new monitoring, new KPI, new SLA, new budget line, new contractual clause, sometimes a new regulatory revision. The cost is not impossibility; it is the institutional debt accumulated for each omitted class.

The upstream perimeter therefore does not only fix what will be seen: it fixes the cost differential between classes seen from the outset and classes re-instrumented after the fact. And that institutional debt, unlike technical debt, is not repaid in a sprint.

The problem is not specific to healthcare. It is isomorphic to difficulties that several mature disciplines have already named: observability in distributed systems, detectability in control theory, support mismatch in statistical learning, causal insufficiency in pharmacovigilance. When the same invariant reappears across four independent formulations, it is not a vocabulary coincidence. It is a structural constraint.

Feedback, and its architecture

The theorem does not imply that the perimeter is frozen. It implies that broadening the taxonomy requires a dedicated procedure, anticipated at design time, without which it does not exist.

That is precisely what regulators have begun to industrialize. The Predetermined Change Control Plan (PCCP), formalized by the FDA in its December 2024 final guidance and progressively adopted for AI devices in 2025 to 2026, is an instrument of explicit feedback:

it obligates the manufacturer to declare in advance the classes of modification achievable without new submission, which amounts to pre-specifying the conversion regime between post-market signal and perimeter revision. The European AI Act, in its articles 9 (risk management system) and 10 (data governance), requires an equivalent device in spirit. The MDR and IVDR complete the architecture through post-market surveillance.

None of these devices eliminates the propagation theorem. They organize its use. They transform the initial blind spot into a declared blind spot and institute a protocol for reducing it over time. A system governed by PCCP or AI Act article 9 is not a system without blind spots. It is a system whose blind spots are contestable. Provided the declaration is honest, and the nomenclature used in the declaration is the one used in monitoring.

The benchmark compiles the risk policy

What the composition of a benchmark actually does deserves naming, and I will do so through one image, a single one, because it is accurate at one point and false everywhere else. The benchmark acts as the implicit compiler of the system's risk policy: it translates a data choice into a behavioral architecture. The image is useful to fix the idea that the dataset is not upstream of the system but inside its conduct. It ceases to hold as soon as it is extended. A compiler produces deterministic execution; a benchmark acts probabilistically, the runtime remains partially adaptive, and the human operator modifies the trajectory. Beyond this image, I will stay with operational terms: visibility perimeter, detectability perimeter, governability perimeter.

The propagation can then be described step by step, and this is where the text ceases to be theoretical. Rare class absent from the benchmark, therefore no class-specific calibration, therefore no dedicated alert threshold, therefore no escalation policy, therefore no targeted monitoring, therefore no associated KPI, therefore no corresponding contractual SLA, therefore no specific budgetary arbitration, therefore no exploitable post-market signal. When the incident occurs, it is interpreted as individual noise rather than as a systemic class.

At each link, nothing malfunctions. Each component does exactly what it was configured to do. The defect is nowhere and everywhere. It is in the initial visibility boundary, which propagated without any engineer ever deciding to ignore the risk.

The three most often forgotten links are precisely the three financial links: KPI, SLA, budgetary arbitration. Without them, propagation remains abstract. With them, it becomes what it is: a capital allocation decision for institutional attention. A CFO who validates a support contract on the basis of an SLA defined without a benchmark coverage declaration signs, without knowing it, an implicit exemption for all out-of-perimeter classes.

The operational bureaucracy

Propagation does not stop at the software. It extends into the organization. It structures escalation workflows, clinical review timelines, support contracts, reimbursement policies, insurance clauses, internal compliance obligations, staffing arbitrations. The benchmark does not only compile a software architecture; it compiles an operational bureaucracy.

The field makes this assertion verifiable.

Consider OCTOPUS, the multicenter observational study on mNSCLC BRAF V600E (n=184, five European countries, survival modeling by SurvTRACE) conducted within the TweenMe framework. This mutation represents approximately 1 to 2 percent of metastatic non-small cell lung cancers. If the validation cohort does not explicitly include this sub-population, the consequence is not that an individual patient will be poorly stratified: it is that the class becomes operationally nonexistent even as it remains clinically real. The presence or absence of a few hundred patients in the validation perimeter can determine, through propagation, the allocation of significant institutional attention over several years of deployment lifecycle.

Visible classes become tracked, funded, auditable, prioritized. Invisible classes become statistically rare, operationally marginal, organizationally silent. The system no longer merely sees certain incidents poorly: it eventually ceases to treat them as central objects of decision.

This is what separates this text from a machine learning safety reflection. The subject is not the robustness of a model; it is the institutional distribution of visibility, of which the model is only the technical core.

The FDA symptom

One fact serves here as a symptom, and it must be handled as such. By early 2026, the FDA had authorized more than 1,400 AI/ML devices, approximately three quarters in radiology, on benchmarks that are not published device by device. The polemical reflex would be to denounce the authorization of black boxes; that is not my point, and it would miss it.

The problem is not that an authorization rests on an incomplete benchmark. Every benchmark is, as conceded from the outset. The problem is more precise: the opacity of the benchmark is not only a methodological opacity; it becomes an opacity on the effective clinical visibility boundary of the authorized system. The market and buyers hold a market authorization certificate; they do not hold the perimeter of the classes the system will know how to escalate, qualify, contest. They purchase a performance without knowing the governability horizon that accompanies it.

The European counterpart is less commented, yet more structuring for continental actors. The AI Act requires documented governance practices and dataset representativeness (article 10), as well as a risk management system operating throughout the lifecycle (article 9), which formally includes perimeter revision in response to post-market signals. On paper, these devices constitute a coherent architecture for the propagation theorem and its conversion regime. In practice, their implementation remains largely to be invented, and French and European industry is playing a part of its institutional credibility there. A representativeness declaration that contents itself with enumerating demographic proportions without declaring the clinical classes covered does not satisfy the propagation theorem. It satisfies its cosmetic version.

The benchmark as upstream promotion gate

This mechanism extends a prior doctrinal line, and I indicate it briefly, because an article should not become dependent on its own corpus to be readable. In the work on the promotion gate, an institutional passage point decides what accesses higher status. The benchmark plays exactly this role for a clinical system. It is its upstream promotion gate.

The principle that follows is simple to state and demanding to hold: a system should not be promoted to production if its validation space does not explicitly cover the classes of incidents it will have to govern, or if it does not declare the classes it does not cover and the conversion regime by which it will subsequently integrate them.

The field gives this rule its substance. The BRAF V600E patient of the OCTOPUS cohort is not a pedagogical example. It is a precise clinical class whose presence or absence in a perimeter decides the system's capacity to recognize it subsequently. PREDICARE, on the territorial decompensation trajectory, ToxTwin on graph-predicted molecular toxicity, raise the same question under distinct forms: what is it, in the perimeter, that guarantees the critical class will be seen, attributed, governed? The answer is never found in the aggregate score. It is found in the composition.

Three clarifications

The thesis does not require an exhaustive benchmark. It requires the declaration of the perimeter and the honest propagation of that taxonomy through to monitoring.

It does not claim that human systems are exempt from the same limit. Human clinical taxonomies have their blind spots; AI changes the scale, speed, and standardization. A clinician locally reconstructs an emerging category; a deployed system propagates its blind spot at the speed of its deployment and with the uniformity of its code.

It does not exempt itself. The identification of critical classes absent from the benchmark depends recursively on a prior visibility structure, that of the designer or the regulator. The

thesis therefore shifts the burden of visibility one notch without claiming to exhaust it. This is a gradient of governability, not a guarantee, and that is exactly what the PCCP and AI Act article 9 regimes are in the process of instituting.

Three requirements for the decision-maker

For a decision-maker, whether buyer, regulator, or medical director, the requirement can be formulated without legal opinion, which is not my object. Three operational demands are sufficient to transform the thesis into an instrument.

First, a coverage declaration by clinical class. For each relevant pathology or sub-population, the supplier produces the number of cases included in the benchmark, the inclusion protocol, and the performance measured on that specific class. The aggregate average becomes one proxy among others, not the primary measure.

Second, an explicit statement of untested classes. This demand is paradoxical in appearance and essential in practice. An organization that knows what it does not know can allocate its attention. An organization that believes it covers everything can prioritize nothing. An honest declaration of blind spots is more protective than an optimistic declaration of total coverage.

Third, a taxonomic propagation clause. The perimeter declared in the validation space appears, class by class, in the monitoring grid, in the alert thresholds, in the tracked KPIs, and in the contractual SLAs. Any discontinuity between these spaces is a governance gap. The minimal control consists of verifying that the nomenclature used to describe the validation space is identical to the one used to describe monitoring. When they diverge, the propagation theorem has already begun its work.

These three requirements are not a metrological ideal. They are the minimal operational translation of the propagation theorem for a buyer who does not wish to sign an implicit exemption.

Conclusion

A clinical benchmark does not only measure a performance. It defines which forms of failure will exist as governable objects for the system, and it fixes the institutional cost of recognizing all the others.

As long as industry ranks its models by average accuracy, it will continue to optimize what it sees and to deploy what it will not know how to govern. Systems will not necessarily become less performant. They will become capable of producing forms of failure they will not know how to recognize, qualify, or contest, and operational responsibility will dissolve exactly where the benchmark left a blank.

The problem is therefore not only statistical. It is institutional.

A broader question surfaces behind this one, and another article will have to address it: if our measurement instruments end up selecting the operationally accessible reality for systems, do we ultimately compress what we hold to be clinically real? It is noted here. It is not unfolded.

Because a benchmark does not only define what a system knows how to do. It defines what it will know how to consider a problem.