Measured Performance, Operational Reliability: The Distinction the Industry Refuses to Make

Why a High AUC Is Neither a Deployment Guarantee nor a Sufficient Measure of Confidence in Regulated AI Systems

Twingital Institute — Jérôme Vetillard — April 2026

Introduction

The public debate on artificial intelligence remains dominated by performance metrics. AUC, F1-score, accuracy, BLEU, perplexity: these indicators structure publications, fuel product announcements, support investor presentations, and occupy a central place in validation dossiers. This centrality is not accidental. Performance metrics have three institutionally powerful properties: they are measurable, comparable, and publishable.

The properties that condition the effective robustness of a system in real deployment situations are, by contrast, far less visible. Probability calibration, the explicit definition of an applicability domain, resistance to distribution shift, the system's capacity to signal its own limits — all of these properties remain, in many contexts, treated as secondary considerations: sometimes important, but rarely structuring.

This asymmetry does not result from ignorance. A large part of the scientific community has long known that a high discriminative metric is insufficient to characterize the safety of use of a model, let alone that of a complete system. The asymmetry is cultural, institutional, and architectural. Performance is favored because it integrates naturally into an economy of comparison. Operational reliability requires reflection on real deployment, on the heterogeneity of inputs, on uncertainty, and on the system's limiting mechanisms. It lends itself far less well to rankings, simplified dashboards, and narratives of linear progress.

The thesis defended here is the following: a model that is performant in the metric sense is not, for that reason alone, a reliable system in the operational sense. These two properties do not derive from one another. They are constructed differently, measured differently, and answer distinct questions. Their confusion constitutes a recurring architectural error in AI systems deployed in regulated environments — not a surface-level engineering problem that an additional monitoring layer could correct.

This thesis has an explicit domain of validity. It applies primarily to systems that produce or inform high-stakes individual decisions: regulatory toxicological prediction, diagnostic decision support, clinical risk scoring, triage, healthcare resource allocation. It does not apply with the same intensity to exploratory systems or low-stakes productivity tools, where errors remain reversible and statistically tolerable.

I. Conceptual Clarification: What Are We Actually Talking About?

The discussion becomes confused when distinct properties are treated as if they lay on a single continuum. It is therefore necessary to distinguish five notions explicitly.

Measured performance refers to a model's ability to produce correct predictions on a given evaluation dataset, according to a data separation protocol and one or more metrics defined in advance. AUC measures the overall discriminative capacity of a classifier across all possible thresholds. The F1-score combines precision and recall for a given threshold. Accuracy/precision expresses the proportion of correct predictions. These metrics are useful. They measure a real property. They do not measure everything that matters in the effective use of a system.

Calibration concerns the relationship between the probabilities or confidence scores produced by the model and the empirically observed frequencies. A well-calibrated model assigning a probability of 0.80 to a set of cases will, on average, see the event occur in approximately 80% of those cases. Calibration is not equivalent to discrimination: two models may have a similar AUC and radically different calibrations. These two properties do not substitute for one another.

Decisional validity refers to the aptitude of an algorithmic output to be used pertinently within a given decision policy. A probability, a score, or a rank only acquires meaning in relation to a threshold, a cost of error, an action timeline, a level of human supervision, and a possibility of complementary verification. The same predictive performance may be decisionally useful in one context and insufficient in another. This level is too often left implicit.

The *applicability domain* (AD) designates the input space for which there are methodologically defensible reasons to consider that the model's predictions remain valid. Every model has, in fact, an implicit AD. The question is not whether it exists, but whether it is explicitly defined, measured, and integrated into the inference pipeline.

Operational reliability refers to the sufficiently predictable, bounded, monitorable, and governable behavior of a system in its real conditions of use. This property does not depend solely on the model. It depends on the complete pipeline: input data, transformations, acceptance rules, out-of-domain case detection, calibration, monitoring, escalation mechanisms, user interaction, and revision timeline.

The central distinction of this article can therefore be formulated rigorously: measured performance characterizes the quality of a model on a given evaluation task; decisional validity concerns how that output can be used in an action policy; operational reliability concerns the behavior of the real system in a deployment environment. These three objects are linked, but must not be conflated.

II. How Performance Absorbed the Very Idea of Rigor

The dominance of performance metrics is not merely a rhetorical effect. It is the product of an institutional history of AI evaluation that deserves to be traced, precisely because it explains the bias's resistance to correction.

The major benchmarks — ImageNet for vision, GLUE and SuperGLUE for language, MoleculeNet for cheminformatics — have played a structuring role in evaluation culture. They enabled inter-team comparison, relative experimental reproducibility, accumulation of results, and identification of real progress. In this sense, their contribution is undeniable. The problem does not lie in their existence. It lies in the progressive transformation of the benchmark into a quasi-substitute for real-world usage.

This drift is reinforced by the logic of leaderboards. A leaderboard rewards what is easy to compare: a final score, obtained under the formal conditions of a task. It values far less the quality of calibration, robustness to distribution shift, the legibility of the applicability domain, or graceful degradation outside the validity space. These properties cannot be summarized on a single line of a table. They have therefore structurally weighed less in the dominant evaluation culture.

To this is added a sequencing bias. In many pipelines, reflection on the system's governability occurs after model training and optimization. One trains first. One evaluates next. One then finally considers calibration, monitoring, atypical case management, and the integration of guardrails. This sequence makes it almost inevitable that reliability is treated as an added layer, when it should be conceived as a constitutive property of the system — precisely the same bias documented for the architectural governance of regulated AI systems, discussed in the previous article of this series.

The result is a collective imbalance. The ecosystem now has very sophisticated tools for comparing measured performance, and comparatively fragmented tools for characterizing real reliability in deployment. This imbalance reflects an implicit definition of what deserves to be optimized.

III. Why Measured Performance Is Insufficient as a Proxy for Reliability

3.1 An Evaluation Protocol Can Honestly Measure the Wrong Thing

The value of a metric depends first on the relevance of the evaluation protocol. A high AUC has no universal significance — it has significance within the framework of the chosen split and distribution.

In cheminformatics, the difference between a random split and a scaffold split is decisive:

- A random split distributes structurally similar compounds between training and test, which favors local interpolation.
- A scaffold split imposes a strict separation between structural families and measures far more the model's capacity to handle genuinely new structures.

Sheridan's (2013) work on QSAR models and the subsequent methodological analyses by Wallach et al. on MoleculeNet showed that such differences in protocol lead to sometimes significant performance gaps — variable depending on the datasets, molecular representations, and models considered.

The general lesson extends beyond cheminformatics. In clinical settings, a random validation can mask temporal or institutional effects that appear as soon as one evaluates prospectively or on an external site. In medical imaging, a naive split may allow spurious correlations linked to equipment or center to persist. A good score on an inadequate protocol does not measure the system's capacity to behave correctly in deployment — it measures its success under the specific conditions of that protocol.

3.2 A Discriminative Metric Does Not Inform Probabilistic Interpretability

AUC is a rank metric: it measures a model's ability to order positive and negative cases, independently of the scale of the scores. It is invariant to monotonic transformations of scores — which is its strength for comparing classifiers, and its radical limitation for judging their decisional reliability.

Two models may have similar AUCs while producing confidence scores of very different magnitude. Yet, in many high-stakes contexts, the user does not only need a good ordering. They need to know whether a score presented as "0.85" corresponds to something stable and interpretable, or whether it is an output useful for ranking but misleading if read as a probability.

A discriminative metric therefore ignores, by construction, several decisive dimensions: probabilistic interpretability, operational prevalence, the asymmetry of error costs, and the local criticality of certain regions of the input space. When the downstream decision rests solely on a relative ranking, this may suffice. As soon as the output feeds an individualized decision, a probabilistic arbitration, a confidence level communicated to an expert, or a partially automated action, it no longer does. A probability displayed to a user is not a simple output format. It is an interpretive commitment.

3.3 The Absence of an Explicit AD Produces Unearned Confidence

By default, a statistical model provides an output for any input compatible with its technical interface. This universal availability of the response must never be confused with a universality of validity.

When no applicability domain is defined or integrated into the pipeline, an input outside the model's validity space receives a score in the same format as a familiar input. For the end user,

the interface does not signal that the prediction was produced in a low-density zone, on an underrepresented type of case, or in a region of the input space where the model has only learned a fragile approximation.

In a toxicological pipeline, this issue becomes acute when a model trained primarily on small organic molecules is confronted with organometallics or coordination complexes whose determining properties are not captured by the featurization employed. The problem is not merely a drop in average performance — it is the silent presentation of a response that retains the appearance of validity.

The benchmark tests the model in its world. Deployment places it in yours. A system without an integrated AD cannot tell the difference.

IV. From Model Quality to System Reliability: A Three-Level Structure

To resolve the confusion between performance and reliability, it is useful to introduce the tripartition established in §I as an operational framework.

The first level is that of *intrinsic model quality*: performance on a task evaluated according to a given protocol, discrimination, possible calibration, robustness.

The second level is that of *decisional validity of the output*: a calibrated probability may be useful for triggering an experimental verification beyond a certain threshold; a poorly calibrated but well-ordered score may remain useful in an exploratory ranking context. This level is too often left implicit — and it is precisely in this implicit space that most misunderstandings between model designers and system users are lodged.

The third level is that of *operational reliability of the system*: management of atypical cases, flow stability, guardrails, legibility of limits, drift detection, temporal behavior, escalation modalities. Reliability is not an isolated attribute of the model. It is a property of the complete system inserted into a real world.

This tripartition avoids a false debate. The problem is not that performance metrics are useless. The problem is that they describe one level, and one level only. The error occurs when they are asked to carry the descriptive burden of the whole.

V. Operational Reliability as a Deliberate Architectural Property

If operational reliability does not derive automatically from performance, it is necessary to specify how it is constructed. In high-stakes systems, it rests at minimum on three architectural decisions made upstream.

First decision: the evaluation protocol must be representative of plausible deployment. This requirement does not imply a single "true" protocol. It implies that the choice of split encodes an honest hypothesis about use. For a system intended to process new chemical families, a scaffold split is more honest than a random split. For a clinical system deployed over time, a prospective temporal validation is often more informative than a random separation. The evaluation protocol is not a technical formality. It is an epistemic commitment about what deployment will look like.

Second decision: calibration must be integrated into the pipeline when the score is used as a probability or interpretable confidence signal. Isotonic regression is suited to data-rich corpora and captures non-parametric relationships between raw scores and empirical frequencies. Platt scaling remains relevant for sparse data or when seeking more stable parametric calibration. The decisive point is not the dogma of the method: it is the integration of calibration as a pipeline component, not as a display option.

Third decision: an operational mechanism for estimating local proximity or validity must be integrated before inference. Methods vary by domain: distance to k nearest neighbors in the feature space, density estimation, ensembles with variance measurement, uncertainty scores, out-of-distribution detection. No mechanism is universal. But in a high-stakes system, the total absence of an explicit mechanism is architecturally indefensible.

These three layers do not substitute for one another. They articulate. A system may be well calibrated within its validity space and remain misleading outside it. It may correctly identify out-of-domain cases while producing poorly adjusted probabilities on accepted cases. It may be evaluated according to a realistic protocol while failing to signal its local uncertainty zones. Operational reliability emerges from their articulation, not from their mere juxtaposition.

VI. The Epistemic Load of Deployment

To make visible a problem dispersed across several sub-literatures, it is useful to have an integrating concept. I propose that of *epistemic load of deployment*.

By this term, I mean the gap between what a model has effectively learned to process under its training conditions, and what real deployment asks it to process — in an evolving, heterogeneous, partially unforeseen, and sometimes out-of-scope usage environment.

This proposal does not claim to identify an entirely new phenomenon. The notions of dataset shift, covariate shift, concept drift, epistemic uncertainty, and out-of-distribution detection already constitute a rich body of technical work. The interest of the proposed concept lies elsewhere: to reunify, in a governance and architecture perspective, phenomena often treated separately that jointly produce the real difficulty of deployment.

The epistemic load of deployment is low in highly controlled environments where the distribution of inputs remains stable and close to the training distribution. It becomes high when

inputs are heterogeneous, practices variable, populations evolving, atypical cases frequent, or error consequences asymmetric and not easily reversible.

This concept invites an inversion of perspective. Instead of asking only whether a model generalizes "well" in the abstract, it leads us to ask what additional burden the real world imposes on this model, and what architectural mechanisms the system puts in place to absorb, signal, or limit that burden. Its operationalization remains an open research program — but its invisibilization in deployment reasoning already constitutes a documentable problem.

VII. What Monitoring Does Not Resolve

Contemporary MLOps platforms — MLflow, Vertex AI Model Monitoring, Amazon SageMaker Clarify — make it possible to track drifts, observe the evolution of feature distributions, trace model versions, and for the most advanced, produce post hoc explanations. These capabilities are useful. They pertain to observability and post-deployment control — not to architectural reliability.

The distinction was developed in the previous article of this series on the architectural governance of regulated AI systems. It applies directly here: a monitoring tool can detect that a distribution is drifting; it does not guarantee that the pipeline refuses or flags cases crossing an uncertainty threshold. A platform can historize aggregated performance; it does not guarantee that the probabilities displayed to the user are correctly calibrated for the decision they must make. Operational reliability is not delegated to monitoring — it is conceived upstream, as a constitutive property of the pipeline.

VIII. Regulatory Frameworks: A Zone of Responsibility Left Open

The frameworks applicable to high-stakes AI systems are evolving rapidly. The MDR's requirements on validation, traceability, risk management, and post-market surveillance are real and substantial. The European AI Act, applicable from August 2, 2026 for the majority of obligations, strengthens requirements on risk management, documentation, logging, and human supervision for high-risk systems.

These texts play an essential role. They create salutary pressure on traceability and accountability. They remain, however, technologically neutral on precise mechanisms: they do not impose any particular calibration method, nor any standard formalization of the applicability domain, nor any measure of the epistemic load of deployment. This neutrality is understandable from a regulatory standpoint. It leaves designers with a strong responsibility: to define themselves what will make their system effectively governable in real situations.

Two systems may both be documented, traced, and monitored, while differing considerably in their capacity to signal their limits before an error materializes. Regulatory compliance is the floor. Operational reliability is the objective.

IX. Implementation Terrain: ToxTwin V2.3+

The principles defended here find an illustrative terrain in the ToxTwin molecular toxicological prediction pipeline, developed within the Twingital Institute. The purpose is not to make this a general proof, but to draw concrete lessons about the feasibility of the reliability layers described.

The audit conducted in early 2026 revealed two structural problems. The first was a circularity in the Ames model validation: the split used in previous versions permitted significant structural similarity leakage. Correcting the protocol — a strict scaffold split with 5 folds on 20,117 compounds — brought the Ames AUC to 0.864 ± 0.056 in cross-validation, compared to non-reproducible values on random split. The second problem was the absence of calibration of output probabilities: the GINEConv OGB model (163 features, dropout 0.3) produced discriminant scores that were not interpretable as empirically faithful probabilities.

The corrections introduced in V2.3 comprise three elements:

1. The introduction of isotonic calibration adjusted on a separate calibration dataset, producing probabilities verifiable on the frozen holdout (SHA256 = 052a2aa2c4cff3d8...).
2. The definition of an operational AD based on the distance to k nearest neighbors in the space of 163 OGB features, with a p95 threshold at 0.332 — a molecule exceeding this threshold receives a negative AD signal before the score is presented.
3. Protocol correction with a frozen holdout guaranteeing the reproducibility of future evaluations.

What this instance proves: that the three layers of operational reliability are implementable in an industrial GNN pipeline, with marginal computational cost. What it does not prove: that the approach is directly transferable without adaptation to other molecular domains. Metal coordination complexes — cisplatin, carboplatin, oxaliplatin — require specialized featurization that the 163 standard OGB features do not support, and constitute the scope of the V3.0 roadmap.

Isotonic calibration is not universally superior for sparse data. The AD threshold retained is not optimal for all use cases.

These results have value as illustrations of architectural feasibility — not as general truth.

X. Limits and Counter-Arguments

A credible position exposes its own limits.

Calibration is not uniformly critical in all contexts. In certain ranking, pre-selection, or exploratory uses, the quality of the ordering may have more practical importance than strict probabilistic interpretability. The argument in this article does not consist in absolutizing calibration, but in recalling that it becomes structuring whenever an output claims to orient an individualized decision under constraint.

The applicability domain is not a methodologically uniform object. The relevant mechanisms change depending on data, models, and domains. There is no universal definition of AD that is directly transposable from cheminformatics to clinical settings, imaging, or LLM-based systems. The general concept remains useful; its implementations must remain specific, cautious, and locally justified.

Aggregated performance can have considerable decisional value even when local legibility remains imperfect. Public health systems or flow optimization systems may benefit from tools for which individual calibration is not the most critical property. The intensity of the thesis defended here depends on the type of decision and the level at which error is judged.

The epistemic load of deployment is an integrating framework, not yet a standardized indicator. It requires operationalization work, methodological comparison, and validation across multiple terrains. Its present interest is analytical: to recall that deployment imposes on the system an additional statistical burden that classical metrics absorb poorly.

Operational reliability does not depend solely on technical properties. It also depends on workflows, user training, supervision modalities, institutional context, and organizational governance. A technically prudent system may remain organizationally dangerous if inserted into a poorly framed use.

Conclusion

The industry optimizes what it measures. Performance has established itself as the dominant language of AI system evaluation because it is measurable, comparable, and publishable. This dominance has produced real progress. It has also sustained a persistent confusion: that which consists in taking the quality of a model on a given protocol as a sufficient approximation of the system's reliability in the real world.

This confusion must be resolved with clarity. Measured performance, decisional validity, and operational reliability are not three degrees of the same property. They are three distinct levels of analysis, constructed differently and answering different questions. A high-stakes system does not become reliable because its AUC is high. It becomes more reliable when it is evaluated according to a protocol coherent with its use, when it renders its outputs interpretable in the right

circumstances, when it delimits its domain of validity, when it knows how to signal extrapolation, and when it integrates these constraints at the very core of its architecture.

The consequence is direct. Operational reliability must no longer be conceived as a late monitoring layer added to an already designed model. It must be conceived as a deliberate architectural property, defined before training, validated in the pipeline, reassessed over time, and articulated to the real decision policy.

What must be measured in regulated systems is not only what is easy to compare. It is what genuinely conditions the possibility of safe, legible, and governable use. The industry should first measure what matters. Then optimize what it measures.

Jérôme Vetillard is VP R&D & Chief Product Officer of Qualees, a digital health company developing TweenMe, a universal digital twin generator for predictive medicine, and leading the Sentinelle IA / PREDICARE program, a territorial predictive medicine platform. He directs the Twingital Institute, an independent think tank dedicated to the industrialization of AI in regulated systems, whose work is published at twingital-ventures.com. Former Head of Technology & Strategy for Healthcare & Life Sciences at Microsoft EMEA, he holds a PhD in Biotechnology from ENS Ulm.