URS GASSER, VIKTOR MAYER-SCHÖNBERGER, AND FABIENNE MARCO

Escaping AI's Measurement Trap

Oversight of artificial intelligence is largely built on benchmarks that fail to capture how models have learned to "perform for the test." A better system would govern through learning and inquiry.

The current approach to artificial intelligence oversight is largely built on measurement. Benchmarks assess capability, red teams probe for failure modes, and evaluation frameworks certify safety and legal compliance before deployment. These instruments can be valuable, but they also share a critical structural vulnerability that AI governance has not yet adequately absorbed: The measurements used to verify AI models are not external to the objects being measured. Rather, this program of measurement operates within the same ecosystem—and is shaped by the same competitive pressures and institutional incentives—that produces the AI technologies it is meant to assess. A result is that the measured behavior of a model may diverge significantly from its behavior when deployed.

Put simply, as AI systems and the organizations building them learn what evaluators look for, the AI model performs for the test, figuring out how to excel in benchmarks without necessarily becoming safer, more useful, or more reliable in real-life scenarios. As evaluation increasingly assesses only the model's capacity to ace that same evaluation, the boundary between system and oversight grows porous.

This kind of situation is known as a measurement trap: When a measure becomes a target, it ceases to be a good measure. Measurement traps are not unique to AI. Standardized testing in education leads schools to "teach to the test" rather than help students develop critical thinking skills. In software development, when productivity is tied to the number of lines of code written, programmers write long, repetitive code, which may or may not be good software. And in business, when bonuses are tied to revenue measures, managers prioritize short-term sales over long-term profitability.

What is unusual about the AI measurement trap is that nothing stands between evaluation and the model's adaptation to game it. Benchmarks become direct inputs to the reward function the model seeks to optimize over time. AI models incorporate the lessons of evaluation while they are being evaluated, and their performance immediately evolves to compensate where necessary. There is no need for a human to understand the testing protocol, design and test adaptations, and then implement them. With the human out of the loop, there is also significantly less likelihood that the adaptation will fail. Humans are less single-minded than machines, and they often stubbornly disregard even obvious optimizations. An AI model won't do that.

In theory, stricter protocols and better benchmarks could mitigate some of the problem, but they cannot eliminate it. That does not make measurement useless; it makes it inherently unstable. The question, then, is not simply how to build better tests but how to design oversight systems—examination protocols, their institutional operators, and relevant legal regimes—that can detect when their own instruments are losing fidelity and reform themselves in response.

Measurement shapes what it measures

In 1975, the economist Charles Goodhart observed what we have been calling the measurement trap: Once actors know how they are being assessed, they begin optimizing for the metric rather than for the underlying objective the metric is supposed to track. First developed in the context of monetary policy, Goodhart's insight has since been applied across economics, management, education, and public administration. Today, the measurement trap is foundational to the sociology of quantification, which demonstrates that metrics are always partial, contestable, and constitutive rather than merely descriptive.

AI governance through measurement has intensified this long-standing dynamic. There is now direct empirical evidence that frontier models can become aware that they are being evaluated. Models also can detect the purpose of the evaluation. Indeed, AI systems make the measurement trap especially difficult to avoid because optimization pressure is woven deeply into how performance is measured, compared, and rewarded. Entire development pipelines are now organized around performance on standardized tests, benchmark suites, and safety audits that are increasingly public, legible, and salient to AI developers angling to outcompete each other.

Given that metrics degrade under optimization pressure, what can be done in response? How can governance systems be built that remain functional when degradation sets in? The standard response has been instrument replacement: When a benchmark saturates, build a better one; when an evaluation regime is gamed, tighten the protocol. That approach has benefits, but it treats each instance of metric failure as a local technical problem rather than a signal of a wider system dynamic. Standard governance is reactive, always trying to catch up with processes it has partly helped to create.

In March of this year, the US National Institute of Standards and Technology (NIST) published a report documenting the consequences of that reactive posture at the system level. Drawing on input from government, industry, and academic experts, the study found that while pre-deployment evaluations have become increasingly sophisticated, post-deployment monitoring remains underdeveloped. Resources and attention continue to be directed far more heavily toward assessment before release than toward system performance in the real world. That imbalance matters because deployed systems interact with evolving and unpredictable technical, institutional, and social environments. In these settings, the report emphasizes, best practices, validated methodologies, and even common terminology for post-deployment monitoring remain nascent. The scarcity of post-deployment evaluation means there is little empirical basis for knowing whether pre-deployment governance instruments retain fidelity once models are operating in the wild, which is the exact condition the measurement trap exploits.

Recognizing that measurement shapes what it measures is the first step toward a new kind of AI governance. In particular, measurement systems themselves should be objects of governance, subject to the same continuous scrutiny as the technologies they are meant to assess. The task is not simply to replace failing instruments more quickly but to build oversight systems in such a way that benchmarking institutions can scrutinize, interpret, and revise their instruments as they operate. In other words, the oversight system should itself be governed through a process of learning from AI models' adaptations to evaluation and from post-deployment assessment of AI performance.

Recent policy frameworks have begun to move in this direction, but only partially. The European Union's (EU) AI Act, for example, does not stop at ex ante assessment to ensure legal conformity. For high-risk AI systems, it also requires serious-incident reporting and post-market surveillance. That is an important institutional advance because it recognizes that compliance at market entry cannot settle the question of safety or legality over the life of a system. But even this policy leaves open the harder question: What should institutions do when the monitoring framework itself becomes predictable, legible to the models it oversees, or weakly connected to the phenomena it is supposed to track?

Design principles for governing by learning

It is not just oversight instruments that inevitably fail to be neutral. It is, more broadly, the oversight framework that cannot be wholly divorced from that which it evaluates. This reality has clear implications for governance.

In particular, it means that governance reform must occur at the institutional level, not just at the level of the benchmark itself. The evaluators of measured systems are part of this measurement ecosystem and must fall under the aegis of a new approach to governance. The same is true of the makers of these systems, as, for instance, the Volkswagen diesel-emissions scandal illustrates. In 2015, the US Environmental Protection Agency learned that the carmaker was deliberately cheating on emissions tests. It did so by targeting the testing regime itself. Volkswagen engineers designed software that determined when an emissions test was underway and altered vehicle performance for the duration of the test. The issue was not simply noncompliance. It was that the company—Volkswagen—adapted to the measurement regime rather than to the regulatory objective behind it.

AI evaluation faces the same structural risk, operating at two levels. Developers can deliberately build systems that behave differently under test conditions than in deployment—the Volkswagen pattern transposed to AI. But the risk runs deeper still. Models do not need to be instructed to game an evaluation. They can optimize for test conditions on their own, without any human directing them to do so. Here, we outline three design principles that can support development of a

more effective AI-oversight architecture. Such an architecture would overcome at least some of the challenges arising from the measurement trap, at the levels of the models themselves, the organizations that build them, and the regulators who supervise the AI industry.

Reflexive oversight. A first-order design requirement is reflexivity: Governance instruments must be built with their own endogeneity in mind. That is, built to turn the lens on themselves, examining not only whether outcomes are being met but whether the categories used to define those outcomes remain fit for purpose.

Consider an AI platform governed by a "user engagement" metric. In it, the AI will learn to prioritize whatever content keeps users scrolling—politically polarizing content, for example—rather than content that might improve the user experience. The maker of this platform should have the capacity and the incentive to recognize that, in optimizing for the desired outcome of engagement, the AI has decided to foster the unwanted intermediate outcome of stoking controversy.

The solution to this sort of problem calls for procedural specification rather than defined metrics. In other words, oversight should prioritize not specific outcomes determined prior to evaluation but rather the means by which decisions are made, who participates in decisionmaking, and what counts as evidence of safety, effectiveness, or harm in light of a particular decision choice. In the case of the hypothetical AI platform, instead of optimizing for user engagement, developers would convene a structured review process—involving platform engineers, independent researchers, and affected user communities—to periodically assess what the system is producing. That process would not ask "is engagement up?" but, rather, who is being harmed, what would count as evidence of failure, and who has standing to raise that question. Governance organized around those questions can adapt when a metric proves inadequate; governance organized around hitting a predetermined number cannot.

Under a reflexive framework, revision of a benchmark is not a concession to failure. It is a necessary acknowledgment that no oversight instrument remains external to the system it governs for long. Reflexivity is already being tested under real circumstances. The US Food and Drug Administration's recently finalized guidance on AI-enabled medical devices requires deployers to submit a Predetermined Change Control Plan—a pre-specified account of what modifications will be made to a system as it evolves in deployment, how those modifications will be validated, and by what means they will be assessed. The evaluation framework is built from the outset to revise itself in response to what the deployed system reveals.

Metrics pluralism. No single metric can capture everything that matters about a complex system. Nor can a unified score, constructed from multiple metrics, accurately account for all components of system performance. AI users therefore benefit from consulting diverse metrics that reflect different elements of a system.

To be clear, we do not mean simply that no metric is perfect. That is a true statement but not an instructive one. Rather, it is that taking the incompleteness of metrics seriously means acknowledging that better design cannot eliminate performance trade-offs. Improving the safety of an AI agent, for instance, may mean sacrificing the transparency of its decisionmaking. Trade-offs are inevitable because gains in one area of performance do not reliably predict performance in another.

As it stands, AI oversight incorporates a great deal of unified scoring: a single score built from multiple benchmarks, intended to capture overall performance. This method is understandable. Regulators want definitive thresholds, policymakers ask for interpretable indicators, and the public wants clear judgments. When organizations choose among AI tools, they want to be able to compare them using similar ratings.

However, unified scoring creates distortion. When a complex system is collapsed into a single score, trade-offs among fundamentally different dimensions of performance disappear from view. Safety, robustness, fairness, interpretability, and downstream effects on humans do not necessarily move together. A model may improve sharply on one axis while degrading on another. A single score hides those tensions and, in doing so, invites overconfidence.

The pressure toward unified scoring is also visible at the regulatory level. Regulators might adopt risk-classification systems that assign technologies to categories triggering different legal obligations, thereby simplifying what is multidimensional and context-dependent. The EU AI Act, for example, adopts a risk-based approach that translates complex assessments of potential harm into administratively legible categories such as prohibited practices, high-risk systems, and transparency obligations. But while this legal architecture makes it easier to handle a problem that would otherwise be unmanageable in scope, it should not be mistaken for providing epistemic completeness.

To overcome this problem of oversimplification, we propose the design principle of metrics pluralism: a deliberate architecture of multiple mutually constraining measures, each capturing one dimension of a complex system rather than claiming to capture the system as a whole. This principle is widely applied outside the realm of AI. For instance, in automated hiring systems, a single "performance" score is insufficient. Instead, such systems deploy a suite of metrics: one for predictive accuracy (whether the system correctly identifies suitable candidates), another for demographic parity (whether selection rates are equal across groups), and a third for stability across different data subsets (whether performance holds when the system encounters applicants

from different contexts or time periods). Because these goals often pull in different directions, the aim is not to eliminate judgment by adding more data, but rather to discipline judgment through structured plurality.

The NIST report is another instance of metrics pluralism in action. Its taxonomy of six monitoring categories—functionality monitoring, operational monitoring, human factors monitoring, security monitoring, compliance monitoring, and large-scale impacts monitoring—is valuable precisely because it resists reducing system performance or social consequence to a single number. Each category tracks a different dimension of what deployed systems do in the world, calling for assessment using distinct methods, forms of evidence, and institutional capacities. Security monitoring, for instance, draws on different expertise and organizational infrastructure than does human factors monitoring, yet both are essential.

Learning from metric decay. Under sustained optimization pressure, stable measurement environments break down. Metrics that initially function as rough but useful proxies lose signal as the systems being measured adapt to, absorb, or route around them. For example, a static benchmark for "AI safety" or "jailbreak resistance" inevitably loses its signal as developers optimize models specifically to pass those tests, or as new attack vectors render old rubrics obsolete. Learning can result from this. The question for governance is not whether breakdown will occur—it will, of necessity—but what degradation reveals and whether institutions are designed to learn from these revelations.

Consider what has happened to benchmarks assessing large language models over the past several years. Benchmarks leaked into training and evaluation pipelines. OpenAI's GPT-4 technical report disclosed that several standard benchmarks had been inadvertently mixed into the training set. The Llama 2 report and subsequent open-source analyses identified similar contamination. The result was not simply a technical irregularity. As public benchmarks became part of the optimization environment, leaderboard gains became harder to interpret. Some reflected genuine capability improvements that could transfer to performance in novel settings, while others captured little more than familiarity with the benchmarks themselves.

The standard response has been to build better benchmarks: harder questions, datasets withheld from public release, and dynamic evaluation sets that update more quickly than training pipelines can absorb them. That response is necessary. But it remains a first-order solution to a second-order problem. It treats contamination as a flaw in particular tests rather than as evidence of the broader structural process described above: Under sufficient optimization pressure, stable measurement environments break down. The training-evaluation boundary collapses. This outcome is inevitable, and it must be treated as such.

Thankfully, the collapse of the training-evaluation boundary is not merely a methodological inconvenience; it is a finding. The breakdown of a formerly useful benchmark reveals how competitive optimization pressure propagates across the training-evaluation divide and discloses an AI development ecosystem that reshapes itself in response to evaluation regimes. By tracking metric degradation, analysts witness the speed with which oversight instruments are absorbed into the very systems they are meant to oversee. In this light, degradation is not just noise. It is data.

This is not to say that governance should tolerate avoidable degradation. The point is that, because decay under optimization pressure is structural rather than accidental, oversight systems can and should be designed to learn from decay rather than merely patch over it. A governance regime that only replaces failed metrics remains reactive, buying a little more time before the next wave of decay. A governance regime that studies how and why metrics fail becomes capable of learning about the ecosystem it is trying to steer.

Creating such a regime requires what management theorist Chris Argyris called *double loop learning*: not simply adjusting behavior within an existing framework when a metric fails but questioning the governing assumptions that made the metric appear adequate in the first place. Single loop learning replaces a saturated benchmark with a better one. Double loop learning asks what that saturation reveals about the governance system and whether assumptions underlying the evaluation regime need revision.

Applied to AI governance, double loop learning means oversight bodies—whether regulators, independent auditors, or standards organizations—should consider adopting a standing monitoring function: systematic review of whether evaluation metrics still correlate with the capacities, risks, or behaviors they were intended to track. Signals to watch for include abrupt benchmark saturation, narrowing variance across high-performing systems, divergence between benchmark gains and real-world outcomes, and evidence of training-evaluation contamination.

And every major evaluation framework should carry a sunset clause, not as an admission of defeat, but as recognition of the structural fact that all metrics decay under sufficient optimization pressure. The issue is not whether a metric will decay, but what its decay reveals, and whether the governance system is designed to recognize decay and learn from it.

From oversight as control to oversight as inquiry

The three principles developed here are not separate recommendations; they address different dimensions of the same structural problem. Together, they express a broader reorientation: from oversight as control to oversight as inquiry.

Governing by learning is importantly distinct from adaptive governance, although the two approaches share many goals. What differentiates them is their design. Adaptive governance

emphasizes continuous adjustment, learning, and resilience in the face of uncertainty. But it still assumes that the instruments generating information about the governed object remain sufficiently reliable to guide that adjustment. The measurement trap reveals why that assumption is untenable in the AI context. Evaluation instruments do not sit outside the system they assess; they are shaped by it and degrade under the optimization pressure they are meant to monitor. Governing by learning therefore takes a different primary object: not the AI system directly, but the integrity of the instruments used to observe it. The question it asks is not "what does the evaluation tell us, and how should we adjust?" but "can we still trust what the evaluation is measuring?"

Overseeing the instruments of oversight requires concrete institutional mechanisms: periodic independent reviews to assess whether metrics still track what they were designed to measure, multiple evaluators with different incentive structures to guard against capture, and sunset clauses that render benchmarks invalid unless periodically revalidated. A regulator overseeing a credit-scoring AI, for instance, might move beyond checking for bias in training data and investigate whether the model has adapted to proxy variables—like zip codes or shopping habits—that circumvent fairness audits. Collectively, such mechanisms are designed to detect signals of strategic adaptation in both AI systems and the organizations deploying them and to route those signals back into governance revision.

One lesson of the EU AI Act is that lifecycle obligations and post-market monitoring are necessary for governance. Another lesson is that post-market monitoring itself must be treated as a potentially decaying instrument rather than a stable window onto reality. Finally, governance by learning demands the involvement of social scientists, psychologists, legal scholars, and affected communities, not just technologists. The most vital questions about AI models concern what they do to people and whose interests the measurement architecture serves. They cannot be answered with technical expertise alone.

These design specifications envision an AI evaluation body whose primary mandate is meta-evaluation rather than direct, object-level assessment. Such a body—whether a regulatory agency, standards organization, academic laboratory, audit firm, or the civil society groups that hold them to account—must treat their own instruments as objects of continuous scrutiny. They must expect to be wrong about their metrics, design for that expectation, and build the capacity to extract learning from moments of wrongness rather than simply replace a failed instrument with a new one.

Existing institutions of AI governance are not currently capable of operating in this way. Fostering that capability will require an interdisciplinary effort. Governance scholars need to identify the specific institutional forms—the bodies, processes, and incentives—that make meta-evaluation viable in practice. On the technical side, observability should be a design requirement from the outset rather than a retrofit. Systems should be built to expose their internal states, decision pathways, and behavioral shifts to external scrutiny, while accounting for the strategic permeability of testing conditions. Simultaneously, social scientists and philosophers of science must examine how measurement constitutes what becomes governable, because in sociotechnical systems, categories do not merely describe behavior but rather actively organize it.

An opportunity for reform

In April 1990, NASA launched the Hubble Space Telescope. The most sophisticated optical instrument ever built, Hubble carried a primary mirror ground to extraordinary precision. But during the telescope's first weeks of operation, it became clear that something was wrong: The images Hubble was capturing were blurred. The cause, it turned out, was not dramatic by engineering standards. The mirror geometry was off by just over two microns, roughly one-fiftieth the width of a human hair.

Such a flaw should not have been possible. The mirror had been tested before launch, and verification instruments confirmed the correctness of its shape. The problem was that those instruments were themselves improperly calibrated, and no effective independent cross-check was used. NASA had built an extraordinarily powerful system for observing distant reality, yet its contractor's tools of measurement distorted what it was seeing.

The repair mission that followed in 1993 is remembered as one of NASA's finest hours. But the key move was not the installation of corrective optics. It was the recognition that the instrument had not merely failed but had been misleading in a consistent way. This finding meant that the distortion itself could reveal what had gone wrong, leading to the conclusion that the verification process itself was to blame.

Today, AI governance faces its own Hubble moment. The problem does not lie in detecting the distortion; it exists and, indeed, is as obvious as a blurry photograph of the heavens. Theory requires that metrics for benchmarking AI models decay, and ample research confirms that they are, in fact, decaying as the models and their creators adapt to oversight. The true problem is that institutions designed to learn from the observed degradation are absent. Before the next generation of oversight tools is constructed, institutions must be built that are capable of recognizing when the tools at hand no longer show the world clearly.

Urs Gasser *is dean of the School of Social Sciences and Technology at the Technical University of Munich.* ***Viktor Mayer-Schönberger*** *is professor of internet governance and regulation at the University of Oxford and coauthor with Gasser of* Guardrails: Guiding Human Decisions in the Age of AI *(Princeton University Press, 2024).* ***Fabienne Marco*** *is a doctoral student at the Technical University of Munich.*