The Open-Box Fallacy: Why AI Deployment Needs a Calibrated Verification Regime

Anonymous Author(s)
Affiliation
Address
email

Abstract

AI deployment in sensitive domains such as health care, credit, employment, and criminal justice is often treated as unsafe to authorize until model internals can be explained. This often leads to an excessive reliance on mechanistic interpretability to address a deployment challenge beyond its intended scope. We argue that the gate should instead be calibrated verification: authorization should be domain-scoped, independently checkable, monitored after release, accountable, contestable, and revocable. The reason is twofold. First, model capability is uneven across nearby tasks, so authorization must attach to a specific use rather than to a model in general. Second, societies have long governed opaque expertise through credentials, monitoring, liability, appeal, and revocation rather than mechanism-level explanation. Recent evidence reinforces this distinction between mechanistic understanding and deployment authority: a 53-percentage-point gap between internal representations and output correction shows that understanding may not translate into action, while one scoping review found that only 9.0% of FDA-approved AI/ML device documents contained a prospective post-market surveillance study. We propose Verification Coverage, a six-component reportable standard with a minimum-composition rule, as the metric that should sit beside capability scores in model cards, leaderboards, and regulatory disclosures.

1 Introduction

AI systems are increasingly moving into operational settings where their outputs can affect access to care, economic opportunity, public services, physical security, or legal standing [U.S. Food and Drug Administration, 2026a, Consumer Financial Protection Bureau, 2022, Wisconsin Supreme Court, 2016]. When these systems are deployed, trust is often sought through greater transparency and explainability [Doshi-Velez and Kim, 2017, Gilpin et al., 2018], with the aim that understanding how a model reaches its output will make deployment decisions more justified, auditable, and safe [Tabassi, 2023, Raji et al., 2020, Mitchell et al., 2019]. This is well motivated: mechanistic interpretability can expose safety-relevant internal structure and reveal failure modes [Lindsey et al., 2025, Hubinger et al., 2024]. However, it does not by itself answer the deployment question.

Systems can be useful in domains where their internal mechanisms remain only partially understood [U.S. Food and Drug Administration, 2026a, De Fauw et al., 2018, Dell'Acqua et al., 2026], and the question facing decision-makers is not whether opacity disappears before use, but what criteria are sufficient for deployment in a specific use case. This is reflected in the EU AI Act [European Union, 2024], which regulates high-risk systems through deployment-facing obligations such as risk management, documentation, human oversight, and post-market monitoring. A credible answer must therefore distinguish model capability and functionality from deployment authority: evidence about how a system solves a task is not yet evidence that its use is appropriately scoped, monitored, and accountable.

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

We introduce the *open-box principle* as the view that mechanistic transparency can provide important evidence for safer deployment, and the *open-box fallacy* as the stronger demand that this evidence serve as the decisive deployment gate. Verification, as we use it, is a deployment-governance regime—formal checks, expert validation, institutional review, contestability, monitoring, and revocation [Tabassi, 2023, Raji et al., 2020, Wachter et al., 2018]—that structures the evidence under which a deployment may be considered; final authority remains with the institutions named in the regime. The unit of authorization is a deployment in a domain, not a model in general: a freely distributed model produces many deployment contexts, each with its own deployer of record.

This position paper argues that verifiers, not interpretability alone, should license AI deployment: mechanistic evidence is valuable, but deployment authority should rest on calibrated verification of a specific use. Mechanistic evidence remains one stream within that regime, alongside behavioural evaluation, independent review, and stakeholder input.

The case for this position is twofold. First, model capability is uneven across nearby tasks, so authorization must attach to a specific use rather than to a model in general [Dell'Acqua et al., 2026]. Second, opacity is not unique to AI deployment: mature institutions have long governed consequential expertise without complete mechanism-level access, relying instead on external checks and revision mechanisms. The question is therefore not whether mechanistic evidence matters, but whether the surrounding regime can make a particular deployment governable.

This paper makes three contributions:

- We introduce the open-box fallacy and explain why it fails: model capability is uneven across nearby tasks, and institutional authorization has often operated under partial opacity.
- We develop calibrated verification as the alternative deployment regime, distinguishing evidence streams, verifier classes, and regime properties.
- We propose Verification Coverage as a six-component reportable standard with a minimum-composition rule, intended to sit beside capability scores in model cards, leaderboards, and regulatory disclosures.

2 The Open-Box Principle and Fallacy

The open-box principle holds that mechanistic evidence is genuinely deployment-relevant: understanding how a system produces its outputs informs whether and how it should be used. Internal inspection of large language models has been shown to recover causal structure—features and circuits that predict behavioural outcomes in ways surface statistics do not capture [Lindsey et al., 2025]. Mechanistic evidence can identify failure modes invisible to benchmark evaluation, distinguish genuine competence from spurious correlation, and inform the design of domain-specific tests.

The principle, correctly understood, supports a bounded inference: mechanistic evidence answers—partially and progressively—the question of how a system works, but not, by itself, whether its use is appropriately scoped, monitored, and accountable in a given context. The first question concerns explanation, the second authorization. The *open-box fallacy* is the stronger inference that mechanistic evidence should be the decisive deployment condition. Mechanistic transparency is neither necessary when outputs can be independently checked, nor sufficient when a deployment lacks domain scope, monitoring, accountability, contestability, or revocation. A system may be internally opaque yet externally verifiable for a narrow deployment; conversely, a system may be mechanistically transparent yet inappropriate for deployment because it lacks those surrounding rails.

2.1 Jagged capability across nearby tasks

A model's capability is not a single property that transfers smoothly from one use to another. Within the same workflow, the same system can improve performance on one task while degrading it on a seemingly similar task. This jagged capability profile is the central empirical fact that makes blanket deployment policy insufficiently granular. Categorical prohibition ignores domains where AI can outperform available alternatives; categorical authorization ignores domains where it remains unreliable. A jagged frontier motivates a granular response: authorization per domain, per verifier class, per use.

The "jagged technological frontier" captures this point: capability can reverse across a task boundary. In a field experiment with consultants, Dell'Acqua et al. [2026] find that AI improves productivity and quality on many consulting tasks, but reduces correctness on a complex managerial task beyond its demonstrated capability. On that task, consultants using AI are 19 percentage points less likely to produce correct solutions than those without AI. Model-level examples tell the same story: Xu and Ma [2025] call this the genius paradox, and Fu et al. [2024] find that models recognize letters but fail to count them reliably, especially when letters repeat.

If AI were uniformly incompetent, deployment debates would be simple. If it were uniformly reliable, trust debates would also be simpler. The practical challenge is that models are capable enough to be useful, opaque enough to be hard to inspect, and uneven enough that neither blanket acceptance nor blanket refusal is adequate. The jagged frontier therefore motivates the paper's positive proposal. Authorization should attach to a deployment context, not to a model.

2.2 Evaluation is already brittle

The same problem appears in evaluation. If authorization should attach to a deployment context rather than to a model, then domain-blind benchmark scores are too coarse to carry the deployment gate by themselves. Frontier models often identify evaluation contexts: Needham et al. [2025] construct a benchmark of evaluation and deployment transcripts and report that Gemini 2.5 Pro reaches an AUC of 0.83 at distinguishing them, against a human baseline of 0.92. van der Weij et al. [2024] show that language models can be prompted or trained to underperform selectively on dangerous-capability evaluations.

Human-feedback evaluation is powerful but bounded: Sharma et al. [2023] show that preference judgments can reward sycophantic responses in some settings, and Wen et al. [2024] show that, on complex QA and programming tasks, RLHF can make incorrect answers more convincing to human evaluators. The April 2025 rollback of a GPT-4o update after sycophantic behaviour reached production illustrates that this is not only a lab issue [OpenAI, 2025]. Chain-of-thought rationales should not be treated as transparent windows either: Chen et al. [2025] find that reasoning models often do not reveal when they used provided hints. Benchmarks and rationales are evidence, not authorization; the next section turns to the institutional conditions that make such evidence deployable.

2.3 Institutional scaffolding under partial understanding

Disciplined use under partial understanding has usually depended on instruments and institutions, not on complete mechanism-level explanation. The closest case is human expertise, but only when the expert is understood as part of an institutional surround rather than as an isolated mind. The analogy is institutional, not anthropomorphic. Societies authorize high-stakes human judgment without transparent access to the mental mechanism that generates each decision, but they do not authorize it blindly. They rely on training, examinations, supervised practice, professional registries, monitoring, liability, appeal, discipline, and revocation.

Authorizing a surgeon to practice does not require a theory of her cognition. It requires evidence of competence within a scoped role, monitoring of outcomes, malpractice exposure, and a procedure for withdrawing authority when warranted. Board exams and professional review may probe domain knowledge directly, including mechanisms of disease or treatment, but the object being authorized is not the neural process that generates each judgment. It is performance under institutional constraints. The institutional regime closes the loop that neither internal inspection nor outcome statistics closes alone. Mechanistic access may improve diagnosis, auditing, and oversight, but it is not the form in which authority is usually granted. Demanding mechanism-level understanding of AI as the universal authorization condition is therefore not a stricter version of standard practice; it is a different standard, applied to a different question.

Prior work notes versions of the asymmetry. Zerilli et al. [2019] argue that transparency demands placed on algorithmic systems can exceed what is demanded of human decision-makers; Kempt et al. [2022] extend the comparison to clinical AI; Jonas and Kording [2017] make vivid that even complete access to a system's parts does not guarantee functional understanding. The conclusion is not that AI should be treated as a human expert. It is that high-stakes authorization should be scoped and institutional rather than model-general and mechanism-dependent.

3 Calibrated Verification

This section turns the negative claim into a positive deployment regime. If authority attaches to a deployment context rather than to a model, the gate must ask whether that context is checkable, accountable, monitored, contestable, and revocable.

The concepts play different roles. Evidence streams describe where deployment-relevant information comes from. Verifier classes describe how outputs are checked. Regime properties describe what a deployment must provide before authority is granted. Verification Coverage, introduced in Section 4, reports whether those properties are present.

3.1 Four evidence streams

Mechanistic evidence. Internal inspection can reveal causal structure that behavioural evaluation misses [Lindsey et al., 2025]. It can sharpen accounts of failure modes, support anomaly detection, and provide distinctive evidence about internal objectives [Hubinger et al., 2024]. But evidence about internals is not the same as deployment authority. In one clinical setting, Basu et al. [2026] report a 53-point gap between detecting hazard information in the model's internal representations and the model's output behaviour. Wu et al. [2025] likewise find that, on steering benchmarks, prompting and finetuning outperform sparse-autoencoder methods. These results do not make mechanistic evidence irrelevant. Instead they show that it must be connected to population validity, post-release stability, accountability, and contestability before it can help authorize deployment.

Behavioural evaluation. Held-out tests, red-team probes, prospective trials, and adversarial elicitation remain central. Scope limits include distribution shift, evaluation awareness [Needham et al., 2025], sandbagging [van der Weij et al., 2024], sycophancy [Sharma et al., 2023, Wen et al., 2024], and unfaithful rationales [Chen et al., 2025]. Behavioural evidence is necessary but not sufficient: a model may pass a benchmark and still drift after release, or pass a benchmark and still fail an affected person's contest.

Independent review. Procedural audit, conflict-of-interest disclosure, and reproducibility checks examine the process that produced the evidence [Raji et al., 2020, Mitchell et al., 2019]. The reviewer may be a public regulator, accredited auditor, standards body, public-interest organization, or independent technical evaluator. Independent review does not substitute for empirical or mechanistic evidence. It checks whether those evidence streams were produced under conditions that resist self-serving framing.

Domain-expert and stakeholder input. Affected parties detect harms, contestability gaps, and context-specific failure modes that evaluators miss. Counterfactual explanation [Wachter et al., 2018] and adverse-action notice show that contestability is achievable under opacity. Stakeholder input should therefore be structured, documented, and tied to review criteria. It is not a substitute for evidence, nor a simple majority vote.

3.2 Verifier classes

Evidence streams are not themselves verifiers. They become deployment-relevant when connected to checks. We distinguish three verifier classes. *Formal verifiers* provide decisive checks inside a specified formal system: proof assistants, type checkers, cryptographic checks, and game rules. *Empirical verifiers* provide probabilistic checks through observation: clinical trials, prospective validation, A/B tests, replication studies, crash statistics, field monitoring, and post-market surveillance. Test suites belong here, not with formal verifiers, since they are practical and high-coverage but incomplete. *Social-normative verifiers* provide institutional checks for decisions that affect rights, opportunity, liberty, or legitimacy: credit, hiring, education, criminal justice, medicine, welfare, and public administration. The classes do not partition the four evidence streams. Each evidence stream can support different verifier classes. Section 4 maps these streams and classes onto the six Verification Coverage components.

Deployment problem	Required regime property
Capability varies across nearby tasks (§2.1)	Domain-scoped authorization
Self-reported evidence may be incomplete or conflicted (§3.1)	Independent checkability
Performance can change after release (§5)	Monitoring
Harm requires a responsible actor (§2.3)	Accountability
Affected parties detect errors evaluators miss (§3.1)	Contestability
New evidence should change deployment status (§3.3)	Revocation

Table 1: Calibrated verification derives its six properties from documented deployment failure modes. Each property answers a distinct failure mode and is non-compensable; strong evidence on one does not offset absence of another in a high-stakes domain.

3.3 Plural verification and deployment accountability

A verification regime is only as strong as its evaluators. Models that recognize evaluation contexts [Needham et al., 2025], sandbag selectively [van der Weij et al., 2024], or sycophantically conform [Sharma et al., 2023] can compromise empirical and human-mediated verifiers, just as they compromise benchmarks; formal verifiers remain decisive only within their specified formal system. The answer is not simply a more transparent model, but a more plural evaluation regime: a structured process in which independent technical evaluators, domain experts, affected stakeholders, and where safe public or open-source auditors evaluate a deployment context under predeclared criteria, with disagreement reported rather than averaged away and with named minority reports preserved when consensus is not reached.

Collective verification should not be understood as simple majoritarian judgment. Three properties make it operationally distinct: evaluators are scoped to the failure modes they are competent to detect; harnesses, rubrics, and failure taxonomies are open while sensitive tests, private data, and dual-use materials are withheld; final responsibility remains with named institutions rather than with the model or the metric. Verification Coverage supplies reportable necessary conditions for deployment; it does not replace institutional judgment about residual risk, public purpose, or domain-specific acceptability. Passing the threshold permits consideration rather than compels deployment, while failure on a required component blocks authorization unless the deployment is redesigned and re-evaluated.

Deployer of record. Authority must attach to a named deployment context, with a primary accountable actor and role-specific duties distributed across the value chain. A freely distributed model therefore produces many separate deployment contexts, not one general authorization. Verification duties travel with the integrator into health care, credit, employment, criminal justice, or other consequential workflows; the same base model may be acceptable for low-stakes drafting and prohibited for autonomous diagnosis or criminal-risk scoring without contestability. Verifiers can still fail; what plural evaluation buys is an auditable trail when they do.

3.4 From failure modes to six properties

The six properties named in the position statement are not arbitrary. Each responds to a deployment failure mode that the preceding sections document, summarized in Table 1.

The properties are not interchangeable. A deployment with strong benchmark performance but no contest path cannot be treated as verified for a rights-affecting domain. A deployment with strong interpretability evidence but no monitoring plan is not adequately verified for a drifting clinical environment. A deployment with full auditing but no revocation trigger leaves no path to act when evidence accumulates. This is why Verification Coverage uses a minimum-composition rule rather than a weighted average.

4 Verification Coverage

A calibrated verification regime needs measurements. The field reports many capability scores, but it lacks a headline report of deployment verifiability. Verification Coverage is a reportable deployment-

level metric in the minimal sense of a structured measurement object: a six-component profile with a minimum-composition rule. It is not a validated universal scalar score. The six components correspond one-to-one to the regime properties:

- **Domain Coverage.** What fraction of real use falls inside the authorized domain? Proxy: log-share of queries inside declared scope.
- **Verifier Strength.** Are outputs formally, empirically, or institutionally checkable, and does evaluation quality hold as the model approaches the overseer's capability [Engels et al., 2025]? Proxy: per-output presence of a check, plus capability-gap-conditioned evaluator-accuracy curves.
- **Monitoring Maturity.** Are post-deployment failures and drift detected? Proxy: presence of a statistically valid surveillance plan in the sense of Dolin et al. [2025].
- **Accountability Clarity.** Is a named actor responsible for the deployment? Proxy: identifiable accountable party in the deployment record.
- **Contestability.** Can affected people obtain reasons and appeal? Proxy: documented contest path with measured response time and outcome distribution.
- **Revocation Readiness.** Are there predeclared triggers and procedures for restricting or withdrawing deployment? Proxy: documented thresholds, incident triggers, named decision-maker, and time-to-action after a threshold breach.

For a deployment d, let $v(d) = (v_1(d), \ldots, v_6(d)) \in \{0,1\}^6$ record whether each component is present and sufficiently documented for the deployment. Verification Coverage is reported first as this six-component profile, which makes explicit which rails of the regime are present and which are missing. The minimum-composition rule treats the weakest required rail as binding: $\mathrm{VC}_{\min}(d) = \min_i v_i(d)$, and a zero on any required rail withholds authorization for that deployment.

A scalar summary $\mathrm{VC}_{w,\mathcal{D}}(d) = \sum_i w_{i,\mathcal{D}}\, v_i(d)$, with non-negative weights summing to one and specified by the relevant evaluator or authorizing institution, may be reported for convenience but must never replace the profile: the same aggregate can hide different verification patterns. Table 2 illustrates this across existing regimes—medicine, credit, employment, autonomous systems, and criminal justice show different verification strengths and gaps across the six rails. Verification Coverage makes those differences reportable rather than collapsing them into model capability or functionality.

5 Existing Domain Patterns and Gaps

Existing high-stakes regimes already distinguish model output from deployment authority. Outputs that are unreviewed and automated, which affect health, safety, livelihood, rights, liberty, or irreversible action, are treated as not sufficient. The relevant instruments differ by domain. Some regimes emphasize authorization and review, others reason and appeal, and others monitoring and incident reporting. The common pattern is not full verification. It is a partial verification with uneven coverage across the six rails.

Scoring rule. Table 2 applies the six Verification Coverage components from Section 4. The coding is diagnostic, not a legal compliance determination. A cell is marked **H** when the component is instantiated by law, regulation, binding agency process, or routine public practice; **M** when it is partial, voluntary, indirect, or unevenly enforced; and **L** when it is thin, practically inaccessible, or not specifically attached to the AI deployment. The table scores the surrounding regime, not the intrinsic quality of any particular model.

Medicine. Medicine has strong domain-scoped authorization and review ability. For example, the FDA clinical decision support guidance asks whether software enables a healthcare professional to independently review the basis for recommendations so the professional does not rely primarily on the software [U.S. Food and Drug Administration, 2026b]. Additionally, FDA maintains a public list of AI-enabled medical devices authorized for marketing, while noting that the list is not comprehensive [U.S. Food and Drug Administration, 2026a]. Clinical AI reporting guidelines such as CONSORT-AI, SPIRIT-AI, and DECIDE-AI likewise treat intended use, human interaction, evaluation context, and error analysis as reportable objects [Liu et al., 2020, Cruz Rivera et al., 2020, Vasey et al., 2022]. The weak rail is post-release monitoring: a 2024 scoping review found that only 62 of 692 FDA-approved AI/ML devices (9.0%) contained a prospective study for post-market surveillance [Muralidharan et al., 2024]; Dolin et al. [2025] use this gap to argue for statistically valid post-deployment monitoring.

Medicine, therefore, shows both sides of the argument: authorization and reviewability can exist under partial mechanistic opacity, but lifecycle monitoring remains incomplete.

Credit. Credit law has comparatively strong instruments of reason-giving and correction. ECOA and Regulation B require specific reasons for adverse action; a 2022 CFPB circular stated that this requirement applies equally to credit decisions using complex and black-box credit algorithms, although the circular was later withdrawn as guidance [Consumer Financial Protection Bureau, 2022, 2025]. The important point is the institutional pattern rather than the circular alone: consequential credit decisions must be tied to reasons, records, and a responsible creditor. Credit therefore scores high on domain coverage, accountability, and contestability, but weaker on AI-specific monitoring and revocation. A denied applicant may receive reasons, but the public usually cannot see whether a deployed credit model is drifting, how often appeals change outcomes, or what event withdraws the system from use.

Employment. Employment has a thinner audit-and-notice pattern. New York City's Local Law 144 prohibits the use of an automated employment decision tool unless it has undergone a recent bias audit, a public summary is available, and required notices have been provided to candidates or employees [New York City Department of Consumer and Worker Protection, 2023]. This is an independent review, but not necessarily individual contestability. A bias audit can report group-level disparity without giving a rejected applicant a meaningful route to challenge the tool's role in the decision or to trigger suspension when a deployment fails. Employment therefore motivates the distinction between audit, notice, contestability, and revocation.

Autonomous systems and critical infrastructure. Autonomous systems show a different profile: stronger incident reporting and monitoring, weaker individual contestability. NHTSA's Standing General Order requires covered manufacturers and operators to report specified crashes involving Automated Driving Systems and Level 2 Advanced Driver Assistance Systems [National Highway Traffic Safety Administration, 2025b]. NHTSA's AV STEP proposal would add a voluntary review and reporting framework for certain ADS-equipped vehicles [National Highway Traffic Safety Administration, 2025a]. The EU AI Act similarly classifies specified AI systems in critical infrastructure, road traffic, employment, credit, education, and law enforcement as high-risk, and requires providers to establish post-market monitoring systems for high-risk AI systems [European Union, 2024]. These instruments check deployed performance without requiring mechanism-level access. The thin rail is contestability for affected non-operators, such as pedestrians, passengers, and communities exposed to system-level risk.

Criminal justice. Criminal justice illustrates low Verification Coverage in a high-stakes setting. In *State v. Loomis*, the Wisconsin Supreme Court allowed consideration of COMPAS with limitations and cautions [Wisconsin Supreme Court, 2016]. Subsequent work found that COMPAS did not outperform a simple two-feature model or crowd-aggregated lay judgments on the Broward County recidivism-prediction task [Dressel and Farid, 2018], and related scholarship has criticized the secrecy and procedural unfairness of COMPAS and similar proprietary recidivism risk tools [Rudin et al., 2020]. Criminal justice may have a named institutional decision-maker, but it lacks strong external verification, monitoring, and contestability when proprietary risk scores enter bail, sentencing, or parole workflows. For Verification Coverage, this is the limiting case: higher stakes require stronger verifiability, not simply stronger explanation demands.

Cross-domain pattern. The table shows why a single capability score cannot answer the deployment question. Medicine has authorization and reviewability but weak surveillance; credit has reasons and correction pathways but weak lifecycle monitoring; employment has audit and notice, but weak revocation; autonomous systems have incident reporting but weak appeal paths for affected non-operators; criminal justice has named institutional authority, but weak verification and contestability. The minimum-composition rule turns this comparison into a diagnosis: $\mathrm{VC}_{\min}(d)$ is set by the weakest rail, and the weakest rail identifies the next repair.

Regime	Domain	Verifier	Monitor.	Account.	Contest.	Revocation
Medicine (FDA CDS; AI/ML devices)	H	H	L	H	M	M
Credit (ECOA/Reg. B; adverse action)	H	M	L	H	H	M
Employment (NYC LL144; state AI laws)	H	M	M	M	M	L
Autonomous systems / critical infrastructure	M	M	H	M	L	M
Criminal justice (post-*Loomis*)	M	L	L	H	L	L

Table 2: Illustrative profile of selected high-stakes regimes on the six Verification Coverage components. Cells use the rule: **H** = component is instantiated by statute, regulation, binding agency process, or routine domain practice; **M** = partial, voluntary, indirect, or unevenly enforced; **L** = thin, practically inaccessible, or not specifically attached to the AI deployment. The table is a diagnostic application of the VC vocabulary, not an empirical measurement or legal compliance determination.

6 Alternative Views

We address seven objections, each either a partial concession, a misreading of the proposal, or an independent reason to accept it.

"General-purpose AI cannot be domain-gated." General-purpose training does not entail general-purpose deployment authority. The jagged capability profile in Section 2.1 makes blanket authorization too coarse in both directions: neither categorical prohibition nor categorical acceptance tracks a frontier that reverses across task boundaries [Dell'Acqua et al., 2026]. Authority attaches to the deployer of record in a specific context; open distribution multiplies deployment contexts, not authorization.

"Some domains have no verifier." That is an intended implication of the framework, not a failure of it. A domain without a plausible verifier class registers a low Verifier Strength score; the minimum-composition rule withholds authorization at that rail regardless of other scores. Decision support under a named accountable human remains available; autonomous consequential use does not.

"Open models make deployment gates unenforceable." Open distribution shifts, rather than eliminates, the locus of duty. Verification obligations attach to whoever integrates a model into a consequential workflow. Enforcement is harder to achieve, but the same artifact produces many distinct deployment contexts, each with its own deployer of record and verification duties. The policy response is to clarify integration-layer accountability, not to treat availability as authorization.

"Interpretability is necessary for detecting deceptive internal objectives." We accept the premise and contest the conclusion. Mechanistic inspection is plausibly the most direct signal for deceptive internal objectives [Hubinger et al., 2024], and the regime weights it accordingly within the Mechanistic Evidence stream. The concession ends there: a 53-point gap between detecting hazard-relevant internal representations and correcting output behaviour [Basu et al., 2026] shows that understanding does not imply control. Mechanistic evidence is necessary where it is the uniquely reliable signal; it is not sufficient as a universal gate.

"Internal auditing is enough" [Raji et al., 2020]. Auditing is necessary but incomplete. It does not specify an external verifier class, an individual contestability path, or a predeclared revocation trigger. The employment row of Table 2 is instructive: NYC Local Law 144 mandates a bias audit yet leaves no meaningful appeal route for a rejected applicant and no revocation condition when disparity persists post-audit [New York City Department of Consumer and Worker Protection, 2023]. Verification Coverage names what auditing leaves open.

"This repackages NIST or the EU AI Act." Existing frameworks establish that AI assurance is multi-dimensional [Tabassi, 2023]; the EU AI Act and FDA guidance already require pieces of monitoring and documentation [European Parliament and Council of the European Union, 2024, U.S. Food and Drug Administration, 2026b]. The contribution is different: naming deployment verifiability as a missing *reportable* object beside capability scores, and imposing a minimum-composition rule

that makes the weakest rail binding. That gap is empirically visible: only nine percent of FDA-registered AI/ML devices include a prospective surveillance plan [Dolin et al., 2025], a deficiency that would directly block authorization under the Monitoring Maturity floor.

"Verification Coverage will create false precision." Reporting the six-component profile alongside any scalar prevents the aggregate from obscuring the weakest rail. The minimum-composition rule is itself anti-precision-inflating: it cannot be gamed by strengthening other components. The remaining safeguards are structural: independent assessment, preserved evaluator disagreement, and revocation when a verifier fails. False precision is an argument for these safeguards, not against measurement.

7 Conclusion

Societies have long authorized opaque expertise through training, credentials, monitoring, liability, appeal, and revocation, and mechanism-level access has often not been the primary basis for authorization. Model capability is uneven enough across nearby tasks that no blanket policy can cohere with that practice. Calibrated verification draws on four evidence streams—mechanistic, behavioural, independent, and stakeholder—and yields six regime properties: domain-scoped, independently checkable, monitored, accountable, contestable, and revocable. Verification Coverage reports those properties beside capability scores under a minimum-composition rule that names the weakest rail rather than averaging it out. The field should report not only what models can do, but also the verification regime that makes a specific deployment governable.

References

Sanjay Basu, Sadiq Y. Patel, Parth Sheth, Bhairavi Muralidharan, Namrata Elamaran, Aakriti Kinra, John Morgan, and Rajaie Batniji. Interpretability without actionability: Mechanistic methods cannot correct language model errors despite near-perfect internal representations. *arXiv preprint arXiv:2603.18353*, 2026.

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, Vlad Mikulik, Samuel R. Bowman, Jan Leike, Jared Kaplan, and Ethan Perez. Reasoning models don't always say what they think. *arXiv preprint arXiv:2505.05410*, 2025. Anthropic Alignment Science Team.

Consumer Financial Protection Bureau. Consumer financial protection circular 2022-03: Adverse action notification requirements in connection with credit decisions based on complex algorithms. CFPB Circular 2022-03, 87 Fed. Reg. 35864, `https://files.consumerfinance.gov/f/documents/cfpb_2022-03_circular_2022-05.pdf`, 2022.

Consumer Financial Protection Bureau. Interpretive rules, policy statements, and advisory opinions; withdrawal. 90 Federal Register 20084, 2025. `https://www.federalregister.gov/documents/2025/05/12/2025-08286/interpretive-rules-policy-statements-and-advisory-opinions-withdrawal`.

Samantha Cruz Rivera, Xiaoxuan Liu, An-Wen Chan, Alastair K. Denniston, Melanie J. Calvert, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: The SPIRIT-AI extension. *Nature Medicine*, 26:1351–1363, 2020. doi: 10.1038/s41591-020-1037-7.

Jeffrey De Fauw, Joseph R. Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine*, 24(9):1342–1350, 2018. doi: 10.1038/s41591-018-0107-6. URL `https://www.nature.com/articles/s41591-018-0107-6`.

Fabrizio Dell'Acqua, Edward McFowland III, Ethan R. Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R. Lakhani. Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. *Organization Science*, 37(2):403–423, 2026. doi: 10.1287/orsc.2025.21838.

Pavel Dolin, Weizhi Li, Gautam Dasarathy, and Visar Berisha. Statistically valid post-deployment monitoring should be standard for AI-based digital health. In *Advances in Neural Information Processing Systems (NeurIPS 2025 Position Paper Track)*, 2025. Argues for statistically valid post-deployment monitoring for AI-based digital health. OpenReview: `https://openreview.net/forum?id=mXBFoHDuil`.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017. URL `https://arxiv.org/abs/1702.08608`.

Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1):eaao5580, 2018. doi: 10.1126/sciadv.aao5580.

Joshua Engels, David D. Baek, Subhash Kantamneni, and Max Tegmark. Scaling laws for scalable oversight. In *Advances in Neural Information Processing Systems*, 2025. NeurIPS 2025 spotlight.

European Parliament and Council of the European Union. Regulation (EU) 2024/1689, annex III: High-risk AI systems. `https://eur-lex.europa.eu/eli/reg/2024/1689/oj`, 2024.

European Union. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence. Official Journal of the European Union, 2024. `https://eur-lex.europa.eu/eli/reg/2024/1689/oj`.

Tairan Fu, Raquel Ferrando, Javier Conde, Carlos Arriaga, and Pedro Reviriego. Why do large language models (LLMs) struggle to count letters? *arXiv preprint arXiv:2412.18626*, 2024.

Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 80–89. IEEE, 2018. doi: 10.1109/DSAA.2018.00018. URL `https://arxiv.org/abs/1806.00069`.

Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M. Ziegler, Tim Maxwell, Newton Cheng, et al. Sleeper agents: Training deceptive LLMs that persist through safety training. *arXiv preprint arXiv:2401.05566*, 2024.

Eric Jonas and Konrad P. Kording. Could a neuroscientist understand a microprocessor? *PLOS Computational Biology*, 13(1):e1005268, 2017. doi: 10.1371/journal.pcbi.1005268.

Hendrik Kempt, Jan-Christoph Heilinger, and Saskia K. Nagel. Relative explainability and double standards in medical decision-making: Should medical AI be subjected to higher standards in medical decision-making than doctors? *Ethics and Information Technology*, 24(2):20, 2022. doi: 10.1007/s10676-022-09646-x.

Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, et al. On the biology of a large language model. *Transformer Circuits Thread*, 2025. `https://transformer-circuits.pub/2025/attribution-graphs/biology.html`.

Xiaoxuan Liu, Samantha Cruz Rivera, David Moher, Melanie J. Calvert, Alastair K. Denniston, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. *Nature Medicine*, 26:1364–1374, 2020. doi: 10.1038/s41591-020-1034-x.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 220–229. ACM, 2019. doi: 10.1145/3287560.3287596.

Vijaytha Muralidharan, Boluwatife Adeleye Adewale, Caroline J. Huang, Mfon Thelma Nta, Peter Oluwaduyilemi Ademiju, Pirunthan Pathmarajah, Man Kien Hang, Oluwafolajimi Adesanya, et al. A scoping review of reporting gaps in FDA-approved AI medical devices. *npj Digital Medicine*, 7(1):273, 2024. doi: 10.1038/s41746-024-01270-x.

National Highway Traffic Safety Administration. ADS-equipped vehicle safety, transparency, and evaluation program (AV STEP). 90 Federal Register 4130, https://www.federalregister.gov/documents/2025/01/15/2024-30854/ads-equipped-vehicle-safety-transparency-and-evaluation-program, 2025a.

National Highway Traffic Safety Administration. Third amended standing general order 2021-01: Incident reporting for automated driving systems and level 2 advanced driver assistance systems. NHTSA Standing General Order on Crash Reporting, 2025b. Effective June 16, 2025. https://www.nhtsa.gov/laws-regulations/standing-general-order-crash-reporting.

Joe Needham, Giles Edkins, Govind Pimpale, Henning Bartsch, and Marius Hobbhahn. Large language models often know when they are being evaluated. *arXiv preprint arXiv:2505.23836*, 2025. Gemini 2.5 Pro reaches AUC 0.83 on evaluation-awareness classification; human baseline 0.92.

New York City Department of Consumer and Worker Protection. Automated employment decision tools. https://www.nyc.gov/site/dca/about/automated-employment-decision-tools.page, 2023.

OpenAI. Sycophancy in GPT-4o: What happened and what we're doing about it. OpenAI blog, 29 April 2025, 2025. https://openai.com/index/sycophancy-in-gpt-4o/. See also the follow-up "Expanding on what we missed with sycophancy" (https://openai.com/index/expanding-on-sycophancy/).

Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 33–44. ACM, 2020. doi: 10.1145/3351095.3372873.

Cynthia Rudin, Caroline Wang, and Beau Coker. The age of secrecy and unfairness in recidivism prediction. *Harvard Data Science Review*, 2(1), 2020. doi: 10.1162/99608f92.6ed64b30.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2023.

Elham Tabassi. AI risk management framework (AI RMF 1.0). Technical Report NIST AI 100-1, National Institute of Standards and Technology, 2023.

U.S. Food and Drug Administration. Artificial intelligence-enabled medical devices. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-enabled-medical-devices, 2026a. Accessed May 7, 2026.

U.S. Food and Drug Administration. Clinical decision support software: Guidance for industry and Food and Drug Administration staff. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-decision-support-software, 2026b.

Teun van der Weij, Felix Hofstätter, Ollie Jaffe, Samuel F. Brown, and Francis Rhys Ward. AI sandbagging: Language models can strategically underperform on evaluations. *arXiv preprint arXiv:2406.07358*, 2024.

Baptiste Vasey, Myura Nagendran, Bruce Campbell, David A. Clifton, Gary S. Collins, Spiros Denaxas, Alastair K. Denniston, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nature Medicine*, 28: 924–933, 2022. doi: 10.1038/s41591-022-01772-9.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2): 841–887, 2018.

Jiaxin Wen, Ruiqi Zhong, Akbir Khan, Ethan Perez, Jacob Steinhardt, Minlie Huang, Samuel R. Bowman, He He, and Shi Feng. Language models learn to mislead humans via RLHF. *arXiv preprint arXiv:2409.12822*, 2024.

Wisconsin Supreme Court. State v. Loomis, 2016 WI 68, 371 Wis. 2d 235, 881 n.w.2d 749 (Wis. 2016). `https://www.wicourts.gov/sc/opinion/DisplayDocument.pdf?content=pdf&seqNo=171690`, 2016.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. AxBench: Steering LLMs? even simple baselines outperform sparse autoencoders. *arXiv preprint arXiv:2501.17148*, 2025. ICML 2025 spotlight.

Nan Xu and Xuezhe Ma. LLM the genius paradox: A linguistic and math expert's struggle with simple word-based counting problems. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3344–3370, 2025. doi: 10.18653/v1/2025.naacl-long.172.

John Zerilli, Alistair Knott, James Maclaurin, and Colin Gavaghan. Transparency in algorithmic and human decision-making: Is there a double standard? *Philosophy & Technology*, 32(4):661–683, 2019. doi: 10.1007/s13347-018-0330-6.