Liberating AI Risk and Harm Evaluations from the Proceduralist Trap: An Evidentiary Rule to Incentivize Internal Stop-Go Commitments

I. Introduction

The continued development and deployment of AI systems which can have significant societal impact has prompted legal regimes across jurisdictions to require AI developers to produce, maintain, and file evidence about their AI development processes, how their AI models behave, and the harms they may cause. This evidentiary turn in AI governance is founded on the sensible view that, at the moment, requiring transparency is the most reasonable way to govern and prepare for the societal impact of AI.[1] The result has been a proliferation of instruments now embedded in laws and voluntary frameworks across the world. Some of the laws in question are designed in a manner that is counterproductive, and this article is focused on showing that and suggesting a corrective path forward.

This article argues that current U.S. state laws and federal and state agency regulations contain a category error. They treat two fundamentally different types of AI evidentiary instruments—those that produce direct evidence about model behavior and harm (which this article terms AI Risk and Harm evaluations) and those that exist to countercheck decision-making systems and verify procedural compliance (which this article terms AI process reviews)—as functionally equivalent. Both are channeled through the same procedural funnel since they mainly need to be performed and recorded, or affected persons notified. This conflation may have been defensible in AI governance's early years, when the priority was understanding the challenges and building evaluative capacity. It is no longer defensible today.

In response, this article proposes that AI developers adopt internal stop-go commitments: ex ante rules that link the results of AIRH evaluations to decisions about whether to pause or halt AI development or deployment. The article then proposes a statutory evidentiary rule to incentivize these commitments. Under this rule, when an AI developer seeks to introduce AIRH evaluations as evidence of due diligence in court, judges would exclude or underweigh such evidence unless the developer demonstrates that it had documented internal stop-go commitments specifying which AIRH evaluation results would trigger pause or halt decisions.

Similar work focusing on bias audits required by the New York Local Law 144 has noted that the audits become ineffective when results are not connected to enforceable obligations and follow-on action.[2] Others argue that audits and impact assessments frequently "stop short" after surfacing disparate impact, and propose embedding model-search and documentation requirements so evaluation results reliably translate into mitigations. This links quantitative

[1] Stephen Casper et al., *Pitfalls of Evidence-Based Policy*, arxiv preprint 12 (Sept. 15, 2025), https://arxiv.org/abs/2502.09618v6; Jakob Mökander, *Auditing of AI: Legal, Ethical and Technical Approaches*, 2 Digit. Soc'y 1, 3–4 (2023).

[2] Marissa Kumar Gerchick *et al*., *Auditing the Audits: Lessons for Algorithmic Accountability from Local Law 144's Bias Audits*, in Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25) 38 (2025).

disparity measurement to procedural or human decision pathways.[3] More in-line with this research, Holden Karnofsky proposes "if-then commitments" that could arguably be equated with the proposed internal stop-go commitments.[4] However, his proposition is not directly linked to the results of AIRH evaluations, while this article's is. The key differences between this literature and the present research is the conceptualization of two different types of evaluations and the intervention of an evidentiary statutory rule that binds evaluation results with stop-go commitments.

The article proceeds in three parts. Part II defines the two categories of AI evidentiary instruments and surveys how laws and policies in the U.S. treat each. It then explains why subjecting AI Risk and Harm evaluations (AIRH evaluations) to purely procedural governance—what this article calls the proceduralist trap—is untenable given what we now know about AI and its risks. Part III introduces internal stop-go commitments, shows how they are distinct from the preparedness frameworks and responsible scaling policies that leading AI developers have adopted, and demonstrates why they are important. Part IV proposes a statutory evidentiary rule to incentivize the implementation of such internal stop-go commitments.

II. AI Risk and Harm Evaluations, Process Reviews, and the Proceduralist Trap

A. Two Categories of AI Evidentiary Instruments

Evidentiary instruments in AI governance can be understood broadly as structured tools and processes that produce evidence about AI development and deployment processes, how AI models behave, and the harms they may cause. Often, AI policy scholars analyze all evidentiary instruments as one category.[5] This article argues, however, that there are two fundamentally different instrument types, distinguished by what they assess and what kind of evidence they produce. The distinction is functional rather than institutional, since a single instrument (such as an algorithmic audit or impact assessment) may contain elements of both. But the distinction matters because the two types serve different purposes and warrant different legal treatment.

1. AI Risk and Harm Evaluations

AIRH evaluations are structured tools and procedures that test how an AI system behaves and what risks or harms it may cause. They probe the model itself—its capabilities, outputs, failure modes, and potential effects[6]—rather than the organizational processes surrounding it. Their ostensible main purpose is to produce direct evidence about potential risk and harm, answering questions like what could go wrong if this system is developed further or deployed?

---

[3] Emily Black et al., The Legal Duty to Search for Less Discriminatory Algorithms, arXiv preprint 11 (June 10, 2024), https://arxiv.org/pdf/2406.06817.
[4] Holden Karnofsky, If-Then Commitments for AI Risk Reduction 1 (Sept. 13, 2024).
[5] *Id*. at 9–10; Stephen Casper et al., *Black-Box Access is Insufficient for Rigorous AI Audits*, in FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency 2254, 2255 (2024).
[6] Laura Weidinger et al., *Sociotechnical Safety Evaluation of Generative AI Systems*, arXiv preprint 6 (Oct. 2023), http://arxiv.org/abs/2310.11986.

AIRH evaluations require threat modeling that considers deployment context.[7] Bias testing examines not just whether outputs differ across demographic groups, but whether those differences would cause cognizable harm in the intended use environment. Similarly, a dangerous capability test for bioweapon synthesis assistance models not just whether the system can provide relevant information, but whether that information, in the hands of plausible threat actors with realistic resources, increases the probability of harm.

Examples of AIRH evaluations include: persuasion tests relevant to manipulation risk;[8] bias, discrimination, and fairness tests examining whether model outputs differ across demographic groups in ways that cause harm;[9] harm-focused components of impact assessments involving the empirical testing of AI system behavior and its effects in realistic deployment conditions;[10] and risk-relevant capability benchmarks measuring metrics like model performance on tasks like scientific reasoning relevant to biological risk and coding ability relevant to cyber risk.[11]

What unites AIRH evaluations is their epistemic function. They produce direct evidence about what an AI system will do in the world,[12] and this kind evidence bears directly on the question of whether development or deployment should proceed. A bias audit that reveals severe disparate impact is not a paperwork problem—it is evidence that deployment will likely violate civil rights. A dangerous capability test that reveals a model can reliably assist with bioweapon synthesis is not a documentation deficiency—it is evidence of foreseeable severe harm. The output of an AIRH evaluation is, in principle, the kind of information that should constrain decisions.

2. AI Process Reviews

AI process reviews are structured examinations of documentation, governance structures, policies, and development processes of AI developers. They verify that prescribed procedures were followed, required documentation exists, and organizational systems are in place.[13] Their ostensible main purpose is to confirm that the AI developer has done what regulators or standards require, not to test what the model will do.

---

[7] *Id*. at 22; Markov Grey and Charbel-Raphaël Segerie, *Safety by Measurement: A Systematic Literature Review of AI Safety Evaluation Methods*, arxiv preprint 16 (June 13, 2025), https://arxiv.org/pdf/2505.05541.

[8] Esin Durmus et al., *Measuring the persuasiveness of language models*, ANTHROPIC, (Apr. 9 2024) https://www.anthropic.com/news/measuring-model-persuasiveness, last visited Jan. 11, 2026; Jérémy Scheurer et al., *Large Language Models can Strategically Deceive their Users when Put Under Pressure*, arxiv preprint 1, 1 (Jul. 2024), https://arxiv.org/pdf/2311.07590; Minqian Liu et al., *LLM Can Be A Dangerous Persuader: Empirical Study of Persuasion Safety in Large Language Models*, arxiv preprint 1, 2 (Apr. 2025), https://arxiv.org/pdf/2504.10430; Grey and Charbel-Raphaël Segerie, *supra* note 4, at 46–47.

[9] Pedro Saleiro, *Aequitas: A Bias and Fairness Audit Toolkit*, arxiv preprint 1, 2 (Apr 2019), https://arxiv.org/pdf/1811.05577; Richard N Landers and Tara S Behrend, *Auditing the AI auditors: A framework for evaluating fairness and bias in high stakes AI predictive models*, 78 AM. PSYCH. 36, 36 (2023).

[10] Bernd Carsten Stahl et al., *A systematic review of artificial intelligence impact assessments*, A.I. REV. 1, 12–13 (2023).

[11] Grey and Charbel-Raphaël Segerie, *supra* note 4, at 41–42; US AISI AND UK AISI, JOINT PRE-DEPLOYMENT TEST: ANTHROPIC'S CLAUDE 3.5 SONNET 5 & 8 (NIST & UK AI Safety Institute, 2024); Nathaniel Li, *The WMDP Benchmark: Measuring and Reducing Malicious Use With Unlearning*, arxiv preprint 1, 2 (May 2024), https://arxiv.org/pdf/2403.03218.

[12] Weidinger et al., *supra* note 3, at 22; Grey and Charbel-Raphaël Segerie, *supra* note 4, at 16.

[13] Jakob Mökander, *Auditing of AI: Legal, Ethical and Technical Approaches*, 2 DIGIT. SOC'Y 1, 3–4 (2023).

AI process reviews answer questions such as: Does the quality management system meet the specified requirements? Is technical documentation complete and accurate? Were the required consultation procedures followed? Do logging and record-keeping systems function as specified?[14] These are important questions but they are categorically different from questions about model behavior and harm.

Examples of AI process reviews include: training data documentation that describes the kind of training data used;[15] process-heavy components of impact assessments and audits focusing on whether required consultation occurred, whether stakeholders were identified, and whether documentation was produced;[16] conformity assessment components that verify whether the provider's quality management system meets regulatory requirements;[17] and reviews of technical documentation for completeness and consistency.[18]

AI process reviews tell us something about how the organization is structured and whether it followed required procedures. This is valuable information since organizational capacity matters, and process compliance can be a proxy for seriousness about AI risks and harms. However, AI process reviews do not directly tell us what an AI system will do in the world. A company can have an impeccable quality management system, complete documentation, and proper governance sign-offs, and still deploy a system that ends up causing serious harm. It is apparent then that AI process reviews and AIRH evaluations answer different questions.

3. The Hybrid Reality and Why the Distinction Still Matters

Many real-world AI evidentiary instruments are hybrids with some components that can be understood as AIRH evaluations and others that are AI process reviews. An algorithmic audit might include both bias testing (AIRH evaluation) and documentation review (AI process review). An impact assessment might include both empirical harm assessment (AIRH evaluation) and verification that stakeholder consultation occurred (AI process review).[19]

The distinction drawn in this article is functional and not institutional since it cuts across existing instrument categories. The distinction matters because the way current U.S. state laws and federal-level policies treat these instruments often tilts heavily toward how it makes sense to treat AI process reviews, overshadowing how AIRH evaluations should be treated. This ends up making AIRH evaluations substantively ineffective.

B. How U.S. State Laws and Agency Regulations Treat Both Categories

---

[14] *Id*. at 15.

[15] Timnit Gebru et al., *Datasheets for Datasets*, 64 Communications of the ACM 86, 86 (2021); Emily M. Bender, Batya Friedman; Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science, 6 Transactions of the Association for Computational Linguistics 587, 587 (2018).

[16] Tjitske Jager and Eric Westhoek, *Keeping control on deep learning image recognition algorithms* in Egon Berghout et al., (eds) Advanced Digital Auditing: Theory and Practice of Auditing Complex Information Systems and Technologies 144–145 (Springer, 2023).

[17] Jakob Mökander et al., *Conformity Assessments and Post-market Monitoring: A Guide to the Role of Auditing in the Proposed European AI Regulation*, 32 Minds and Machines, 241, 260 (2022).

[18] Francesco Sovrano et al., *Simplifying Software Compliance: AI Technologies in Drafting Technical Documentation for the AI Act*, 30 Empirical Software Engineering 1, 2 (2025).

[19] Mökander *supra* note 10, at 14–15.

The U.S. lacks comprehensive federal AI legislation comparable to the EU AI Act. In the U.S. AI governance currently occurs through a patchwork of sector-specific rules, state legislation, and voluntary frameworks.[20] However, an examination of key enacted measures—including NYC Local Law 144, NY RAISE Act, California's AB 2013, Colorado SB 24-205, California's TFAIA (SB-53) and Colorado SB 21-169—alongside agency rules such as Colorado Insurance Reg. 10-1-1, California Civil Rights Council Regulations, HTI-1 Rules and BIS Proposed Rules reveals a consistent pattern: each requires only that the AIRH evaluation or AI process review is conducted, with no clear mechanism connecting findings to development or deployment decisions.

NYC Local Law 144 illustrates how law treats AIRH evaluations. The law requires employers to conduct a bias audit of automated employment decision tools, publish results, and notify candidates.[21] On that issue, that is all it provides for, so the provision functions as a transparency requirement with no decisional connective consequence.

The New York RAISE Act, 2025, similarly emphasizes disclosure over enforcement for AIRH evaluations as it requires large developers of frontier models to implement and publish a safety and security protocol addressing catastrophic risks and to report significant incidents to the state within 72 hours.[22] However, it sets no specific substantive threshold that would bar deployment nor suspension of models if the assessments reveal extreme risks safe for the vague characterization 'unreasonable risk of critical harm'.[23]

Notably, it also empowers the Attorney General to institute a civil action as against a developer not only if they fail to report, but also if they fail to adhere to their frontier AI framework.[24] This penalty is triggered by non-compliance with a self-authored document of the AI company, not by what the evaluation findings reveal. Crucially, this provision could be interpreted as an indirect form of the substantive accountability this article proposes below, but its force is entirely contingent on what the developer chose to put in the framework—which the law leaves entirely open. The gap this creates is precisely what the evidentiary rule proposed in Part IV is designed to fill.

California's AB 2013 illustrates the same pattern for AI process reviews. The law requires developers of generative AI systems to post documentation about their training data.[25] This is purely a disclosure obligation—nothing conditions deployment on the substance of what is disclosed.

Colorado SB 24-205 combines both instrument types (AIRH evaluations and AI process reviews) but does not clearly connect either to decisions. Although it does require impact assessments alongside a general duty of reasonable care, these are legally drafted as parallel obligations.[26] The impact assessment produces a document while the reasonable care duty says manage the risks. They are not connected in the legal text. This point is especially clear when

---

[20] Tatevik Davtyan, *The U.S. Approach to AI Regulation: Federal Laws, Policies, and Strategies Explained*, 16 J. L. Tech. & Internet 223, 225 (2025).

[21] N.Y.C. Admin. Code § 20-871 (2021); NYC Department of Consumer and Worker Protection (DCWP), *Automated Employment Decision Tools: Frequently Asked Questions* 2 (June 29, 2023). See also, Lucas Wright et al., *Null Compliance: NYC Local Law 144 and the Challenges of Algorithm Accountability*, in *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '24) 1, 3(2024).

[22] Responsible AI Safety and Education (RAISE) Act, S. 6953-B, 2025 Leg., Reg. Sess. § 1421(1)(A) (N.Y. 2025) (codified at N.Y. Gen. Bus. Law art. 44-B).

[23] *Id.* at § 1421(2).

[24] *Id.* at § 1427.

[25] Cal. Civ. Code § 3111 (added by A.B. 2013, ch. 817, 2024 Cal. Stat.).

[26] Colo. S.B. 24-205, 74th Gen. Assemb., Reg. Sess. § 6-1-1703 (3) (Colo. 2024).

you consider the drafting next to that used in Article 9 of the EU AI Act,[27] which connects AIRH evaluations to the duty to mitigate and remediate based on the risks unearthed.

In addition, unlike NYC Local Law 144 and NY RAISE Act, Colorado SB 24-205 goes beyond purely procedural regimes by requiring deployers not only to conduct impact assessments but also to describe how they will mitigate risks identified through those assessments.[28] Even so, its mitigation duty remains general, because the statute does not tie particular findings, such as foreseeable algorithmic discrimination, to any mandatory pause, halt, or other binding deployment consequence.

Even California's SB-53, which explicitly targets the most severe risk from frontier AI models, follows this structure. The law requires large frontier developers to publish a "frontier AI framework" describing how they assess whether their models pose catastrophic risks and how they mitigate such risks. It mandates that the developers in question document and disclose their approach but does not specify what evaluation findings must trigger pause or halt decisions.[29] The framework is a transparency instrument rather than a decisional constraint. Similar to Colorado SB 24-205, this law includes partial progress towards substantive regulation by requiring developers to apply risk mitigations.[30] However, this mitigation requirement remains open-ended, because it does not specify what degree of assessed risk must trigger a pause or halt in development or deployment, leaving compliance to depend largely on whatever mitigation the developer considers appropriate.

Colorado SB 21-169 contains an AIRH evaluation component that requires insurers to provide an assessments of the results of a risk-management framework that determines whether their models and data sources unfairly discriminate based on certain characteristics.[31] Nonetheless, it ultimately remains trapped in the same proceduralist logic identified here, because the statute pairs that evaluative demand with reporting, monitoring, and attestation obligations which qualify as AI process reviews without linking adverse findings to any commitments for continued use or deployment.

Under agency regulations, this conflation persists. Colorado Insurance Regulation 10-1-1 (which is the implementing regulation of Colorado SB 21-169 discussed above) requires testing to detect discriminatory outcomes in insurance practices, focusing on whether a system's real-world operation produces legal harm alongside periodic reporting and officer attestation which are qualities of both types of evaluations discussed here.[32] Similar to Colorado SB 24-205 it also invokes mitigation. However, that obligation remains general and a risk-management posture rather than tied to specific evaluation findings in a way that would legally require a particular development or deployment response.

The federal agency rules in Health Data, Technology, and Interoperability (HTI-1) and the Bureau of Industry and Security's (BIS) proposed rule require process reviews. HTI-1 includes governance and documentation obligations that function AI process reviews. The rule inquires whether internal controls exist, whether specified policies have been implemented, and whether

[27] Regulation (EU) 2024/1689, art. 9, on the Artificial Intelligence Act, O.J. (L 2024/1689) (EU).
[28] *Id*. at § 6-1-1703(3)(b)(II).
[29] Cal. Sen. B. 53, 2025–2026 Reg. Sess. § 22757.12. (Cal. 2025).
[30] *Id*. at § 22757.12. (a)(3).
[31] Colo. Rev. Stat. § 10-3-1104.9(3)(b)(III)–(V) (2024).
[32] 3 Colo. Code Regs. 702-10-1-1-5 (2025).

documentation about a predictive intervention has been produced and maintained.[33] Similarly, BIS requires covered developers to disclose the results of AI red-team testing in specified safety-critical domains (but does not require that they are done).[34] It also requires that they report any mitigation measures taken in response to the test results.[35] Tellingly, any mitigation language in these regimes is generally not tied to the process-review requirements in a clear, instrument-specific way. Instead, mitigation tends to remain at the level of broad risk-management or nondiscrimination commitments, allowing the law to focus on whether governance documentation exists and is complete rather than on whether review findings legally require a concrete change in development or deployment.

The survey thus reveals a consistent structural feature. Whether the requirement concerns an AIRH evaluation or an AI process review, and whether it appears in statute or rule, current U.S. laws and agency regulations generally treat both as procedural issues. This procedural equivalence, despite the functional difference between the two instrument types, is the structural feature that the article focuses on.

The table below illustrates how U.S. AI governance instruments—whether requiring direct evidence of model behavior and harm (AIRH Evaluations) or verification of procedural compliance (AI Process Reviews)—channel both through identical procedural frameworks. The critical gap: no instrument clearly specifies what evaluation findings should trigger pause or halt decisions.

Law / Regulation	Instrument Type	What the Law Requires	How Results Appear to Be Treated
NYC Local Law 144 (2021)	AIRH Evaluation	Annual independent bias audit; publication of results on company website; candidate notification	No standard for audit outcomes; no remediation mandate. Enforcement guidance confirms employers are not required to take any specific actions based on results, even if significant disparate impact is found.
NY RAISE Act (2025)	AIRH Evaluation	Safety and security protocol addressing catastrophic risks; incident reporting within 72 hours; documentation of capability and risk evaluations; civil action for failure to adhere to the frontier framework	No substantive threshold that would bar deployment. Law does not require developers to achieve any particular risk level or suspend models even if assessments reveal extreme risks. Penalties apply only to reporting failures.
Colorado SB 24-205 (AIRH component)	AIRH Evaluation	Impact assessment evaluating risk of algorithmic discrimination prior to deployment and annually thereafter	Requires some form of risk mitigation based on results. However, law does not appear to specify what findings would require halting deployment. Compliance creates rebuttable presumption of reasonable care.
California SB 53 / TFAIA (AIRH component)	AIRH Evaluation	Catastrophic risk assessments; confidential submission of summaries to state agency	Requires mitigations to be applied based on assessment results. However, law does not

[33] *Health Data, Technology, and Interoperability: Certification Program Updates, Algorithm Transparency, and Information Sharing*, 89 Fed. Reg. 1192, 1272 (Jan. 9, 2024); 45 C.F.R. §§ 170.315(b)(11)(vi)(C) & 170.315(b)(11)(v)(A)(1).

[34] Establishment of Reporting Requirements for the Development of Advanced Artificial Intelligence Models and Computing Clusters, 89 Fed. Reg. 73612, 73614, 73617 (proposed Sept. 11, 2024).

[35] *Id.* at 73617.

			appear to specify what evaluation findings must trigger pause or halt decisions.
Colorado SB 21-169/Colorado Insurance Reg. 10-1-1 (AIRH component)	AIRH Evaluation	Testing aimed at detecting discriminatory outcomes in insurance practices	Mitigation language exists but is not tethered to the AIRH evaluation as such. Remediation is framed as part of broader governance posture, not as defined consequence following from particular evaluation findings.
Colorado SB 24-205 (Process Review component)	AI Process Review	Documentation of principles, processes, and personnel used in risk evaluation; description of data categories processed	Results treated procedurally. Compliance appears to be satisfied when documentation exists and is complete.
California SB 53 / TFAIA (Process Review component)	AI Process Review	Internal governance practices ensuring implementation of Frontier AI framework	Results treated procedurally. Compliance appears to focus on whether governance practices are in place.
California AB 2013	AI Process Review	Documentation about training data to be posted publicly	Purely a disclosure obligation. Nothing conditions deployment on the substance of what is disclosed.
HTI-1 (Federal Agency Rule)	AI Process Review	Governance and documentation obligations for predictive decision support interventions	Verification mechanisms asking whether internal controls exist, policies have been implemented, and documentation has been produced and maintained.
Colorado SB 21-169/Colorado Insurance Reg. 10-1-1 (Process Review component)	AI Process Review	Periodic reporting; officer attestation	Process-review structures designed to confirm that prescribed procedures and documentation are in place.
California Civil Rights Council Regulations	AI Process Review	Retention and recordkeeping duties for employment-related automated decision systems	Creates audit trail for later review. Structures compliance around traceability and documentation rather than direct evidence about model behavior.
BIS Proposed Rule (Federal)	Both AIRH Evaluation and AI Process Review	Disclosure of AI red-team testing results in safety-critical domains; reporting of mitigation measures taken	Requires transmission of evaluation results but document does not indicate that specific findings trigger mandatory halt or pause decisions.
Key Finding: Across the surveyed U.S. laws, both AIRH evaluations and AI process reviews are generally treated as procedural requirements. Compliance appears to be satisfied when evaluations are conducted and documentation is complete. Even where mitigation language exists, it typically does not operate as a binding consequence keyed to specified evaluation results.			

C. The Proceduralist Trap

As demonstrated above, several U.S. laws and policies require that AIRH evaluations be conducted and sometimes reported but do not prescribe how the results should inform AI development or deployment decisions. That is what this article calls the proceduralist trap. Existing scholarship has documented the proceduralist tendencies of AI governance[36] and

[36] Monika Zalnieriute, *Against Procedural Fetishism in the Automated State*, in Zofia Bednarz and Monika Zalnieriute (eds) Money, Power, and AI: Automated Banks and Automated States 225 (Cambridge University Press, 2023); Monika Zalnieriute, *Against Procedural Fetishism: A Call for a New Digital Constitution*, 30(2) Indiana J. Global Leg. St. 227, 228 (2023).

warned of audit washing,[37] but this article offers a different diagnosis. It argues that some proceduralism is appropriate and the problem is not that AI governance is too procedural. The problem is that current law contains a category error, treating two functionally distinct instrument types similarly.

1. The Category Error

AI process reviews verify that prescribed procedures were followed, and confirm that records are being generated, retained and filed. Consequently, their purpose is inherently procedural,[38] and treating them procedurally makes sense. AIRH evaluations exist for a different purpose. They probe whether the AI system poses risks or will cause harms, asking questions that are categorically different from those asked by AI process reviews. Yet current U.S laws and policies channel AIRH evaluations through the same procedural funnel as AI process reviews. This is the category error that this article is centered on.

2. Consequences of the Category Error

The proceduralist trap helps explain why audit washing emerges from current regulatory design. When law requires only that AIRH evaluations be conducted and not that their findings inform decisions in specified ways, AI developers face an incentive structure that rewards performative compliance. The cheapest way to satisfy a procedural requirement is to conduct an AIRH evaluation that checks the required box. If law does not require that findings trigger specific consequences, it is a rational response for AI developers to underinvest in AIRH evaluations.

The proceduralist trap also results in AIRH evaluations conferring legitimacy where none may be warranted. The existence of AIRH evaluation programs signals to regulators, investors, and the public that AI systems have been vetted. Observers may reasonably assume that responsible AI developers would not proceed after discovering serious risks. But when law does not require that findings constrain decisions, this assumption may be unwarranted. Finally, the most alarming consequence could eventually be the development and deployment of AI that eventually results in significant societal harm.

3. Why the Conflation Is No Longer Defensible

Early on, this proceduralist approach to AIRH evaluations as appropriate for a nascent AI governance field made sense. In the 2010s, a plausible argument could be made that little was known about AI and its risks and hence a proceduralist approach that prioritized transparency and information-gathering was fitting.[39] Indeed, early-stage governance often prioritizes information-gathering over substantive constraint, particularly when the phenomena being governed are poorly understood.

Whatever the merits of this justification in earlier years, two reasons have weakened it considerably. First, although it is still dogged by issues, the AIRH evaluation infrastructure has

---

[37] Ellen P. Goodman and Julia Tréhu, *AI Audit-Washing and Accountability*, German Marshall Fund, 7–10 (Nov. 2022).
[38] Mökander *supra* note 10, at 15.
[39] Andrew Selbst, *Institutional View of Algorithmic Impact Assessments*, 35 HARV. TECH. L. J. 117, 121 & 123–24. (2021).

matured. For example, bias auditing has become a recognized practice with established methodologies, while AI Safety Institutes conduct respectable safety evaluations.

Moreover, AIRH evaluations have produced decision-relevant findings. Bias audits have revealed discriminatory patterns in deployed systems while safety evaluations have documented capabilities in AI systems that warrant serious attention. Overall, many harms from AI systems are no longer hypothetical, and some harms have occurred in contexts where AI developers had conducted evaluations and satisfied applicable procedural requirements. Against this backdrop, continuing to treat AIRH evaluations procedurally is difficult to justify.

III. The Importance of Internal Stop-Go Commitments

A. Understanding Internal Stop-Go Commitments

This article uses the term internal stop-go commitments to mean ex ante rules adopted by an AI developer which clarify the identifiable metrics for each AIRH evaluation that will lead the AI developer to pause further development or withhold deployment of an AI model. The commitments must be established before the relevant AIRH evaluation results are known and must be reasonably specific—identifying particular metrics, thresholds, or outcomes that trigger a stop decision, such that an observer can determine whether the condition was met.[40] General statements of principle that do not indicate how particular AIRH evaluation results will influence development or deployment decisions do not fit this definition.

The term applies only in scenarios where the AI developer conducts the AIRH evaluations itself, commissions it from a third party, or conducts it in close collaboration with a third party. The term covers situations in which an AI developer has no high-level policy guiding AIRH evaluations, and situations where an AI developer has adopted a high-level responsible scaling policy or preparedness framework that contemplates a hard stop in development or deployment. In those cases, internal stop-go commitments are ex ante rules that identify which AIRH evaluation outcomes count as meeting the stated stop condition. Good faith implementation of internal stop-go commitments would ensure that AIRH evaluations are not sidelined as passive documentation exercises.

B. Industry Practice and the Gap from Internal Stop-Go Commitments

Leading AI developers take a variety of approaches to deciding what their acceptable risk thresholds are, and what they ought to do when those thresholds are crossed. Some of these approaches closely mirror internal stop-go commitments as defined here, but none exactly matches the definition. A survey of industry practice reveals two main categories.

The first category consists of what we refer to as *quasi stop-go commitments*: conditional pause language with significant managerial discretion. Under this approach, AI developers include

[40] This is close to the "if-then" commitments proposed by Holden Karnofsky, see, Carnegie Endowment, Holden Karnofsky, *If-Then Commitments for AI Risk Reduction*, 1, 1 (Sept. 13, 2024). Notably, "if-then" terminology was borrowed from, Yoshua Bengio *et al*., *Managing extreme AI risks amid rapid progress* 384 SCIENCE 1, 5 (2024) (author's version).

stop-go language in their risk preparedness policies but with broad qualifiers and a focus on catastrophic risks. Meta and Microsoft, for instance, both state that they will stop or pause frontier development if risks reach a critical level and cannot be mitigated.[41] Although this is close to internal stop-go commitments, the operative hedge—"cannot be mitigated"—still preserves wide discretion over how they judge whether further mitigation is possible, thereby failing the requirement that each decision-path be determined ex ante.

OpenAI's Preparedness Framework defines high and critical risk categories across domains such as biological threats, cybersecurity, and persuasion, and pairs them with safeguards.[42] However, it does not disclose any invariant stop triggers. Recent updates of the Preparedness Framework even say OpenAI may adjust safety requirements if rivals deploy high-risk systems, leaving commitments explicitly contingent on competitive context.[43] This preserves discretion over how safeguards are judged satisfied. Similarly, Google DeepMind's Frontier Safety Framework introduces Critical Capability Levels with mitigation menus for a raft of risks,[44] yet no public hard stop is specified at any Critical Capability Level.

Anthropic's Responsible Scaling Policy comes closest to this article's definition of internal stop-go commitments but still contains a significant number of qualifiers. The company commits to pause scaling or delay deployment whenever capability outpaces safety procedures, and to withhold deployment at high levels of risk until thorough red teaming finds no risk of catastrophe.[45] However, several triggers remain loosely specified and the policy operates at a high level of abstraction. Furthermore, all the frameworks discussed above do not cover societal impact risks such as disinformation and discrimination, even though these are categories of harm that some of these companies already perform AIRH evaluations for.

The second category consists of *compliance-oriented commitments*. AI developers in this category have adopted internal commitments that only follow the requirements of regulatory frameworks and soft law such as the EU AI Act and th GPAI Code of Conduct.[46] It is important to note that AI developers generally have strong incentives to publicize governance mechanisms that are robust. The absence of visible disclosure can therefore be informative and it can be reasonably concluded that where credible internal stop-go commitments exist, there would be at

[41] Meta, *Frontier AI Framework* Version 1.1, 1, 4 (Feb. 3, 2025); Microsoft, *Frontier Governance Framework*, 1, 5 (Feb. 2025).
[42] OpenAI, *Preparedness Framework Version 2*, 1, 4–6, 10–12 (Apr. 15, 2025).
[43] Kyle Wiggers, *OpenAI may 'adjust' its safeguards if rivals release 'high-risk'*, TechCrunch (Apr. 15, 2025), https://techcrunch.com/2025/04/15/openai-says-it-may-adjust-its-safety-requirements-if-a-rival-lab-releases-high-risk-ai, last visited Dec. 23 2025.
[44] Google Deepmind, *Frontier Safety Framework Version 3.0*, 1, 4 (Sept. 22, 2025); Four Flynn, Helen King and Anca Dragan, *Strengthening our frontier safety framework*, Google Deepmind (Sept. 22 ,2025) https://deepmind.google/discover/blog/strengthening-our-frontier-safety-framework/, last visited Dec. 23 2025.
[45] Anthropic, *Responsible Scaling Policy Version 2.2*, 1, 11 (May 14, 2025).
[46] Foo Yun Chee & Supantha Mukherjee, *Code of practice to help firms comply with AI rules may apply end 2025, EU says*, Reuters (July 3, 2025), https://www.reuters.com/business/media-telecom/code-practice-help-companies-with-ai-rules-may-come-end-2025-eu-says-2025-07-03/, last visited Dec. 22 2025.

least some form of disclosure.[47] Conversely, the lack of any public disclosure indicates that such commitments likely do not exist or exist only in a minimal and non-operational form.

Three Approaches to AIRH Evaluation Results

Comparing what this article proposes, what current law requires, and what leading AI developers do

Internal Stop-Go Commitments What This Article Proposes	**Current Law** US, EU & UK Requirements	**Industry Practice** What Leading AI Developers Do
1. **Ex Ante Specification** Rules established before AIRH evaluation results are known 2. **Specific Triggers** Defined thresholds linking particular AIRH evaluation results to stop decisions 3. **Binding Constraint** Halt or pause is mandatory when threshold is crossed 4. **Limited Discretion** No managerial override once criteria are triggered	1. **Conduct AIRH Evaluations** Require AIRH evaluations to be performed 2. **Document Results** File AIRH evaluation findings and maintain records 3. **"Manage" Risks** Reduce risks to self-declared "acceptable" level 4. **Proceed** No required stop conditions tied to specific AIRH evaluation findings	1. **Qualified Language** *"If risks cannot be mitigated, we will stop"* 2. **Preserved Discretion** Management decides what AIRH evaluation results count as "mitigable" 3. **Competitive Contingency** Safety requirements may "adjust" based on rival behavior 4. **Narrow Scope** AIRH evaluations for catastrophic risks only; societal harms excluded
AIRH evaluation result ***determines*** decision	AIRH evaluation result ***informs*** process	AIRH evaluation result ***enters*** discretion

Figure: Comparison of approaches to linking AIRH evaluation results to development/deployment decisions

C. How Internal Stop-Go Commitments Transform AIRH Evaluations Into Procedural Safeguards

As demonstrated in Part I, the laws and policies around current AIRH evaluations primarily enforce procedural diligence in AI development. Internal stop-go commitments can help counter this turn. When internal stop-go commitments are implemented, AI governance can pursue substantive goals without waiting for external enforcement. An AIRH evaluation generates information about a model's risk profile while its attendant internal stop-go commitment supplies an actionable rule for responding to that information. Together, they create a form of built-in

[47] Collins G. Ntim, *Governance structures, voluntary disclosures and public confidence: The case of UK higher education institutions*, 30(1) Accounting, Auditing & Accountability J. 65, 66–67 (2017); Onipe Adabenege Yahaya, *The Influence of Environmental Disclosure on Corporate Information Asymmetry* 51 Accounting Rev. 217, 217–218 (2025).

accountability: AI developers commit in advance to act on the evidence by pausing when danger signs appear.

D. The Example of Operational Safety in Aviation

The proposed internal stop-go commitment is not a theoretical abstraction invented for the AI sector. Similar ideas are being used in other risk-heavy industries. The most salient example can be found in commercial aviation, where airlines operate under pre-defined protocols that mirror the exact structure of the internal stop-go commitments proposed here.

In the daily operation of a commercial airline, safety is not left to ad hoc judgment calls made under commercial pressure. Instead, airlines operate within a framework where specific risk indicators trigger mandatory operational halts. This framework is centered on the concept of an *Acceptable Level of Safety Performance*—a concrete, ex ante definition of the maximum risk the organization is willing to tolerate.[48] Airlines typically operationalize this boundary through rigorous risk matrices that combine the severity of a potential hazard with its likelihood.[49] A high-risk finding functions as an automatic red light and the flight is grounded. The decision is pre-determined by the matrix.

This operational dynamic is practically identical to the internal stop-go commitments proposed in this article. In both cases, the decision to halt operations is made before the specific incident occurs, based on criteria determined in advance. The evaluation result leads directly and automatically to a decision, bypassing executive discretion and insulating the safety decision from immediate commercial incentives. The sophisticated aviation safety infrastructure is mandated by law, specifically the FAA's Safety Management System Regulations.[50] Instructively, the FAA is not required to write the pre-flight checklists for airlines. Instead, the Regulation mandates only that airlines must have these kinds of measures and follow them. While such direct legal mandates are one way to compel AI developers to put in place internal stop-go commitments, there exists a more calculated way to achieve the same results, which Part IV proposes.

IV. An Evidentiary Rule to Incentivize Internal Stop-Go Commitments

A. Design of the Proposed Evidentiary Rule

Parts II explained how AIRH evaluations, as implemented in U.S. laws and policies, falls into a proceduralist trap. Part III argued that internal stop-go commitments can play a crucial role in correcting this development. This Part suggests a statutory evidentiary rule to incentivize internal stop-go commitments. In essence, the rule would condition the admissibility and weight of a AI developer's evidence that it performed AIRH evaluations on the existence of contemporaneous internal stop-go commitments.

---

[48] Safety Management System for Certificated Airports, 75 Fed. Reg. 62,0013-62,0015 (Oct. 7, 2010).
[49] Federal Aviation Administration, Advisory Circular No. 120-92D, *Safety Management Systems for Aviation Service Providers* 3-32 (May 21, 2024).
[50] 14 C.F.R. Part 5.

The basic form of the proposal is a statutory rule conditioning the admissibility and evidentiary weight of AIRH evaluations on the existence of internal stop-go commitments. The rule would be triggered when AI developers rely on AIRH evaluations as evidence of due diligence in litigation. As analyzed in II.A.3 above, most instruments are hybrids. Thus, this rule must be applied to AIRH evaluations components rather than entire documents. If an AI developer attempted to introduce the fact that they conducted AIRH evaluations as evidence of due diligence, the court would exclude or significantly underweigh such evidence unless the developer demonstrates that, at the time the AIRH evaluations were designed and conducted, it had documented internal stop-go commitments. Furthermore, the rule requires a nexus between the AIRH evaluations and the set commitments. These commitments must have specified exactly which AIRH evaluation results would trigger a mandatory pause or halt in development or deployment. Courts will also have the power to scrutinize whether stop-go commitments themselves were reasonable.

A possible formulation could be:

> In any civil action involving an AI system, no evaluation, audit, or assessment of AI risks or harms prepared by a developer or deployer of the AI system shall be admissible as evidence on the issue of that party's lack of fault, due care, or compliance with law unless it is shown that at the time of preparing the evaluation the party had established specific written criteria or commitments that would trigger defined remedial actions or halting of deployment in response to the evaluation's findings.

The evidentiary rule would have an asymmetric application, as the condition would apply only when AIRH evaluations are invoked by defendants as a shield to establish due diligence. The rule would not restrict the use of AIRH evaluations as a sword against such actors.[51] In addition, the rule would not mandate public disclosure of internal stop-go commitments. AI developers would remain free to treat such commitments as confidential information,[52] and the rule would intervene only at the evidentiary stage, when AIRH evaluations are affirmatively submitted as proof of due diligence.

B. Evidentiary Parallels to This Rule

The statutory rule proposed above mirrors other existing legal evidentiary rules. Under the U.S. Federal Rules of Evidence, business records can be admissible despite their out-of-court nature because they possess special indicators of reliability. The business records exception codified in Rule 803 (6) allows a record to be admitted if it was prepared "at or near the time" of the event by someone with knowledge and was kept in the ordinary course of a regularly conducted activity.[53] Importantly, this exception only applies if the circumstances indicate sufficient trustworthiness. If the record appears unreliable, stale, or made for litigation purposes, a court

---

[51] *Dillenbeck v. City of Los Angeles*, 69 Cal. 2d 472 (Cal. 1968) (it was held that internal safety rules are admissible both to evidence the applicable standard of care and to show a failure to meet that standard when violated); RESTATEMENT (THIRD) OF TORTS: LIABILITY FOR PHYSICAL AND EMOTIONAL HARM § 13 cmt. a & d. (Am. L. Inst. 2006).

[52] Ulla-Maija Mylly, *Transparent AI? Navigating Between Rules on Trade Secrets and Access to Information*, 54 INT. REV. INTELLECT. PROP. COMPET. LAW 1013, 1026–27 (2023).

[53] Fed. R. Evid. 803(6)(A)-(C); Erik C. Olson**,** *Issues concerning the Admissibility in Federal Courts of Business Records Containing Opinions or Diagnoses under Federal Rule of Evidence 803(6)*, 4 HASTINGS BUS. L.J. 153, 155 (2008).

may reject it even if it technically fits within the exception. Another example can be located in the so-called Daubert Rule. In *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, the U.S. Supreme Court held that courts should scrutinize the methodology and reasoning underlying the expert's opinion before admitting any expert evidence.[54]

Both the business records exception and the Daubert Rule are directly in parallel with the conditioning logic underlying this article's proposed rule. Just as the admissibility of business record depends on procedural reliability, so should an AIRH evaluation only count as reliable evidence of due diligence if it is produced pursuant to a credible, pre-existing governance process. And just as courts exclude junk science that lacks methodological rigor,[55] they should exclude junk AI governance—AIRH evaluations that are untethered to meaningful internal processes and that were not designed to influence decision-making.

C. Why AI Developers Will Rely on AIRH Evaluations as Evidence During Litigation

Some may claim that the proposed rule would be ineffective because AI developers are unlikely to rely on AIRH evaluations as evidence of due diligence during litigation. Examples from analogous fields show that this assumption would be incorrect. From the historical trajectory of environmental law, once environmental impact assessments became ubiquitous, firms routinely introduced them to prove they had considered risks, even if the project proceeded unchanged.[56] The same pattern appears in product liability cases involving pharmaceutical and automotive sectors.[57]

Within the AI landscape, the primacy that several laws and policies place on conducting AIRH evaluations reflects a growing industrial best practice that AI companies will likely use as evidence of reasonableness and due care. Furthermore, many leading AI developers have already adopted Responsible Scaling Policies and Preparedness Frameworks, positioning their internal evaluation frameworks as evidence of diligence. The proposed evidentiary rule would anticipate this trend and take strategic advantage of it.

D. Strategic Advantages of the Proposed Evidentiary Rule

The proposed rule directly responds to the proceduralist trap by removing the legal upside of proceduralist AIRH evaluations. Under current conditions, a company might conduct a perfunctory civil rights impact assessment and later proclaim that the assessment was a fulfillment of their duty, hoping that the existence of a review will impress regulators or judges. The proposed rule nullifies that ploy; a purely procedural review cannot be used as a defense unless it had substantive teeth. And given that regulators are increasingly mandating these

---

[54] Daubert v. Merrell Dow Pharm., Inc., 509 U.S. 579, 589–95 (1993).

[55] *Daubert,* 509 U.S. at 600–01 (Rehnquist, C.J., concurring in part); Kumho Tire Co. v. Carmichael, 119 S. Ct. 1167, 1179 (1999); Margaret A. Berger, The Supreme Court's Trilogy on the Admissibility of Expert Testimony, in Reference Manual on Scientific Evidence 28 (2d ed. 2000 & 3d ed. 2011) (Federal Judicial Center/National Academies Press).

[56] Jane Holder, Environmental Assessment: The Regulation of Decision Making 10, 44 (Oxford University Press, 2004).

[57] Restatement (Third) of Torts: Prod. Liab. § 4 cmt. e (Am. L. Inst. 1998) (for mandated regulatory standards); Maggie Wittlin, *Evidence of Compliance*, 74 DePaul L. Rev. 827, 832 (2025).

evaluations, not doing them is often not an option, so doing them properly becomes the rational choice.

Another significant advantage of this proposal is that it leverages the existing litigation and liability system rather than requiring the development of a new federal AI agency.[58] It is fundamentally an evidentiary addition that operates within traditional tort, contract, or administrative lawsuits. This is efficient and pragmatic. Tweaking evidence law can often be achieved at the state level by legislatures or court rule committees, and it slots into ongoing cases.[59] Every time there is an AI-related harm being litigated, this rule could come into play without a new enforcement bureaucracy. Private litigants essentially become adjunct enforcers of AI governance, as they have incentive to probe what internal processes the company followed or failed to follow.

Substantively, the rule is agnostic about what the acceptable threshold or mitigation is. That preserves flexibility for AI developers to innovate and to determine what level of risk is acceptable or how to measure risk. This is important given the diversity of AI applications. A healthcare AI tool might measure risk in terms of false diagnosis rates; a self-driving car in terms of disengagements per thousand miles; and a hiring algorithm in terms of adverse impact ratio. The AI company is free to choose how safe is safe enough—but having chosen, it is held to that choice under the glare of litigation. Finally, good faith on the part of AI developers (with regards to the proposed rule) can also be policed comfortably. The adversarial litigation process can unearth truth through evidence and cross-examination and traditional tools like depositions, subpoenas, and requests for production can thus become the machinery that would enforce the rule's spirit.[60]

---

[58] Davtyan *supra* note 17, at 227–30.

[59] Rules Enabling Act 28 U.S.C. § 2071-77 & 2074(b) (2000); Michael Teter, *Acts of Emotion: Analyzing Congressional Involvement in the Federal Rules of Evidence*, 58 CATH. U. L. REV. 153, 154 (2008); *Symposium, The Politics of[Evidence] Rulemaking*, 53 HASTINGS L.J. 733, 739 (2002) (comments of Judge Fem Smith); G. Alexander Nunn, *The Living Rules of Evidence*, 170 U. PA. L. REV. 937, 938 (2022).

[60] FED. R. CIV. P. 27.

Appendix

Each of the ten laws and regulations below conditions compliance on performing, documenting, or disclosing an AI Risk and Harm evaluation or AI Process Review — not on halting or modifying deployment based on what that evaluation reveals. Where deployment-related duties exist, they employ undefined open-textured standards such as "unreasonable risk," or "reasonable care" or are structurally severed from the evaluation findings themselves.

NYC Local Law 144 (2021)
§ 20-871(a)

"[U]nlawful...to use an automated employment decision tool...unless: 1. Such tool has been the subject of a bias audit...and 2. A summary of the results...has been made publicly available." Conditions use on audit being conducted and published not on outcome. No passing threshold, no discontinuation requirement if disparate impact is found.
NY RAISE Act (2025)

§§ 1421(1)(a),(d); 1421(2); 1427
"Implement a written safety and security protocol"; "record...and retain...information on the specific tests and test results." § 1421(2) bars deployment only if it "would create an unreasonable risk of critical harm" — undefined. § 1427 penalises failure to comply with developer's own framework.
Pre-deployment duties are documentary. The only deployment bar uses an undefined open-textured standard. § 1427 enforcement is tethered to self-authored process documents, not evaluation findings. A developer may publish a purely procedural framework with no stop-go commitments and face no consequence.
California AB 2013 (2024)

§ 3111(a)
"[T]he developer...shall post on the developer's internet website documentation regarding the data used...to train the generative artificial intelligence system or service."
Pure transparency rule. Twelve disclosure items are required but no disclosed fact triggers any consequence beyond posting — no halt to training, no dataset alteration, no enforcement mechanism in the operative text.
Colorado SB 24-205 (2024)

§§ 6-1-1703(1),(3)(a),(3)(b)(II); 6-1-1706(3),(5)
Impact assessment must include "an analysis of whether the deployment...poses any known or reasonably foreseeable risks...and, if so...the steps that have been taken to mitigate the risks." Compliance creates a rebuttable presumption of reasonable care (§ 6-1-1706(5)).
Mitigation is a documentary duty — record what was done, not achieve any outcome. Completing the paperwork insulates against liability; the substantive content of findings creates no binding consequence.
California SB 53 / TFAIA (2025)

§§ 22757.12(a)(2),(3),(d)
Framework must address "[d]efining and assessing thresholds" for catastrophic risk and "[a]pplying mitigations...based on the results of assessments." Quarterly transmission of risk-assessment summaries to Office of Emergency Services required.
No provision bars deployment based on assessment findings. The framework mandates a process for defining thresholds and applying mitigations — not a prescribed threshold or mandated outcome. An unfavourable assessment is not itself a violation.
Colorado SB 21-169 (2021)

§§ 10-3-1104.9(3)(a),(b),(d)

Commissioner rules "establish means by which an insurer may demonstrate...that it has tested" for unfair discrimination. § 10-3-1104.9(3)(d): documents disclosed are "confidential and privileged...not subject to discovery or admissible in evidence in any private civil action." Testing is permissive, not prescriptive — rules establish means to demonstrate compliance, not binding test standards. Crucially, § 10-3-1104.9(3)(d) makes the documentary record of testing inadmissible in private litigation, affirmatively shielding the results from legal scrutiny.
Colorado Insurance Reg. 10-1-1 (2023/2025)

Section 5 (Governance & Risk Management Framework)
Insurers must document: testing methodology, assumptions, results, and "steps taken to address unfairly discriminatory outcomes"; ongoing performance monitoring; annual framework reviews.
Requires documentation of steps taken — not that any particular remediation standard be met. Nothing in the regulation bars use of a model that produces an unfavorable test result; the only halt trigger is the underlying statutory prohibition on unfair discrimination itself, not the test output.
HTI-1 Federal Rule (2024)

§§ 170.315(b)(11)(v)(A)(1); (b)(11)(vi)(A)–(C)
Predictive DSIs "must be subject to analysis of potential risks and adverse impacts" (risk analysis); "practices to mitigate risks" (risk mitigation); and "policies and implemented controls for governance" (governance).
No fairness or accuracy threshold is specified for certification. The rule does not direct decertification or removal based on disclosed source attributes or adverse risk-analysis findings. Governance and documentation are mandated; substantive remediation based on findings is not.
BIS Proposed Rule (2024)*
Proposed § 702.X(b)(iii); 89 Fed. Reg. 73614, 73617
"[T]he results of any developed dual-use foundation model's performance in relevant AI red-team testing, including a description of any associated measures the company has taken to meet safety objectives."
Pure quarterly information-collection regime. Does not prohibit deployment based on red-team findings, establishes no passing standard, and does not authorize BIS to enjoin a development run because of disclosed flaws. (Note: rule was not finalised before the change in Administration.)
CA FEHA ADS Regulations (2025)

§§ 11009(f); 11013(c)
"[R]elevant to any such claim or available defense is evidence, or the lack of evidence, of anti-bias testing or similar proactive efforts...including the quality, efficacy, recency, and scope of such effort, the results of such testing or other effort, and the response to the results." Records retained four years.
No rebuttable presumption is created. Anti-bias testing is merely relevant evidence — two-sided, not a shield. Evidentiary weight depends on quality and the employer's response to findings. No prescribed test, no passing ratio, no discontinuation requirement for an ADS exhibiting disparate impact.

E. Limitations and Residual Concerns

In many realms, and especially in AI governance, no policy intervention is a panacea, and this proposal is no exception. The proposed evidentiary rule would only come into play when there is a legal dispute that reaches a court or similar tribunal. Some may claim that this inherently limits its direct effect in situations where an AI harm results in litigation. It would also be less effective in jurisdictions with weak tort enforcement or limited discovery rights. However, the rule's influence does not require frequent courtroom invocation. Its presence in the legal landscape can alter corporate behavior, as AI developers might become wary of what could happen if they were to be sued.

Another significant limitation is the fact that this rule does not directly prohibit any particular risky AI development or deployment. It acts as a mechanism to make the AIRH evaluations layer of AI governance actually functional. The trade-off is that it cannot eliminate many bad outcomes—it can only reduce them by pushing companies to respond to what their assessments find. Despite these limitations, the rule is justified as a forward-looking measure because it aligns with the direction in which AI governance is heading.

V. Conclusion

While the evidentiary rule proposed in this article will not prevent all AI harms, it is a targeted, principled intervention at the critical juncture of risk evaluation and decision. It transforms AIRH evaluations from mere procedural tools into substantive safeguards. Its limitations are real, but they are outweighed by the rule's potential to drive a positive feedback loop: companies internalize stronger practices to avoid liability, which reduces harms and builds trust, which in turn lessens the need for more heavy-handed regulation.

In the grand scheme, internal stop-go commitments allow AIRH evaluations to function as meaningful tools for legal and ethical control of AI development and deployment. The proposed evidentiary rule operationalizes that linkage in a manner that is both innovative and grounded in long-standing U.S. evidence law principles and doctrine, offering a promising path to more responsible and accountable AI development and deployment in the years ahead.

GENERATIVE AI USAGE STATEMENT

We disclose that Generative AI has only been used to edit and improve the grammar and fluency of this article.

REFERENCES

** The BIS Proposed Rule (89 Fed. Reg. 73612) was not finalised before the change in Administration and does not currently impose binding obligations.*