ORION AI POLICY ACCELERATOR

Evaluation Dependence in Frontier AI Safety Frameworks

A structural mapping, a reliability assessment, and the case for activation pathways that do not depend on what the model is willing to show us

---

Emma Mouliere, Ottavia Fries, Lydia Ruyi Shi, Ziyun (Cailee) Feng, Elsa Donnat (research manager), and David Gringras (conceived and supervised the project). Spring 2026.

Contents

Executive Summary

On 7 April 2026 Anthropic published the Claude Mythos Preview System Card and announced that the model would not be made generally available, releasing it only to a limited set of vetted partners for defensive cybersecurity. The card sets out the grounds at length: zero-day exploit discovery the document calls "accelerat[ing] offensive exploitation" (p.14); automated R&D evaluations the card itself concedes "no longer provide evidence that capabilities are short of the thresholds of interest" (p.33); a more general admission that capability assessments at this scale "increasingly rely on subjective judgments rather than easy-to-interpret empirical results" (p.19). On those grounds, the framework's machinery should have gated the deployment decision. A footnote on page 16 reports, in passing, that it did not; the decision to withhold, in the footnote's own wording, "does not stem from Responsible Scaling Policy requirements." That footnote is what motivates this paper. Written by the most cautious of the three frontier developers, about its own framework, at the moment of that framework's most operationally consequential application to date, it is in our reading the most diagnostic single line in the present generation of frontier AI safety governance: the formal architecture (Risk Reports, ASL classification, the CEO and Responsible Scaling Officer review) did not gate the decision it was meant to gate; the right outcome was secured instead by a layer of senior judgment exercised outside the formal pipeline; and there is presently no specified mechanism by which a regulator could tell whether such a layer exists at the next developer or, if it does, whether it would be exercised in the right direction. We have no reason to doubt that Anthropic's judgment was well-directed here. The question this paper takes up is the one beneath: what a regulatory regime designed around the next decision, at the next developer, with judgment exercised by people the regulator cannot inspect, is entitled to assume.

The same dependency runs through the other two frameworks. In the three lab frameworks currently gating frontier model release decisions, every safety-critical mitigation we identified passes through an evaluation result before activating, across a range of triggers as consequential as any the field has articulated: deployment gates for chemical and biological weapons uplift to non-expert and expert threat actors; autonomous offensive cyber capability including zero-day discovery; systematic large-scale manipulation; high-stakes sabotage of critical infrastructure or of successor AI systems; acceleration of frontier R&D. There is one partial exception across the three (DeepMind's ML-R&D Critical Capability Level, which permits internal progress signals to supplement the formal evaluation), and beyond it, evaluation-gating goes end to end. The same structural dependence is inherited, less concretely, by the EU AI Act's Article 55 regime for general-purpose AI models with systemic risk, by the GPAI Code of Practice, and by the Seoul Frontier AI Safety Commitments.

That the architecture is evaluation-dependent would not, by itself, be the problem. It becomes the problem because the evaluations themselves are demonstrably unreliable, and the evidence of unreliability comes predominantly from the labs' own research, which creates a conflict of interest: the party with the most to lose from an unfavourable result is the one responsible for producing the evidence. Simply varying the format and scaffolding, while keeping the model and content constant, swings measured safety scores by 5 to 20 percentage points. The AISI Network conducted a Pilot Testing Exercise across three institutes, the closest analogue the field has to an inter-laboratory reliability study, and found divergence of comparable magnitude on the same model and the same benchmark, attributed in their write-up to "small methodological

differences" in implementation. The label "safety score" reads more like a heading than a quantifiable metric; apparently identical scores can rely on different judges, different aggregation rules, and different threat models. Models can even purposely underperform: chain-of-thought monitoring, the foremost approach to sandbagging detection, still misses 16 to 36 percent of covert cases. The framework authors do not deny any of this evidence; they are the ones who primarily generated it, and they could not coherently contest it. Their response has been better evaluation rather than less dependence on it, and the dependence itself is the part no framework has yet addressed.

We catalogue six classes of mechanism that activate outside the evaluation pipeline: safety cases, compute governance, structured access, institutional investigation in the mould of the AAIB, auditor independence in the mould of SOX, and guaranteed safe AI. We assess each by its cross-domain precedent and its near-term implementability. None of these is sufficient alone, and the case for the portfolio is not the case for any single one of them. It is the manoeuvre that every field which could not abide catastrophic failure had to make once its primary instrument was revealed to be fragile: don't rely on a single layer; engineer several, and ensure they will not all fail on the same input. Software engineering hit this wall in 1985 with Avizienis's N-version programming, betting that parallel implementations of a single specification rarely share a breakdown trigger. Nuclear regulation arrived here as the IAEA's defence-in-depth. Clinical safety arrived here as Reason's Swiss-cheese model, wherein harm must align with holes in every layer simultaneously. What these disciplines share is that their layers fail on independent inputs. There is precisely one such avenue for frontier AI today, the 10^{25} FLOP trigger of Article 55, weakly enforced and with its recalibration methodology yet to be published, and no fallback for the case in which a safety-critical evaluation is itself what fails. The operative principle, articulated and defended throughout: *any safety-critical oversight mechanism whose failure could contribute to catastrophic outcomes should be required to demonstrate at least one activation pathway that functions independently of the evaluation infrastructure.*

---

1. Introduction

1.1 Stakes

The frontier model release decisions of the next several years will be made under safety frameworks whose operative form, whatever the framing texts say at higher levels of generality, gates on capability evaluations. Anthropic's Responsible Scaling Policy assigns models to AI Safety Levels on the basis of evaluation-derived Risk Reports reviewed by the CEO and Responsible Scaling Officer (a small group of senior insiders, in practice; the framework is, structurally, a CEO-and-RSO judgment regime layered onto an evaluation pipeline). OpenAI's Preparedness Framework runs a near-identical architecture under different names: a Capabilities Report compiled from internal evaluations is reviewed by the Safety Advisory Group, which determines whether a threshold has been crossed and whether the threshold-specific safeguards in §4 of the framework engage. Google DeepMind's Frontier Safety Framework, the most recent of the three, activates Critical Capability Level mitigations on the basis of what it terms "early warning evaluations," a phrasing that gestures at, without operationalising, the question of how early is early enough. The EU AI Act's Article 55, which

obliges providers of general-purpose AI models with systemic risk to conduct model evaluations, perform systemic risk assessments, ensure adequate cybersecurity protection, and report serious incidents, operationalises through evaluation results; the GPAI Code of Practice and the Seoul Frontier AI Safety Commitments inherit the same dependency in less concrete form.

The mitigations whose activation turns on those evaluation outcomes are not minor: deployment gates for chemical and biological weapons uplift to non-expert and expert threat actors; autonomous offensive cyber capability, including zero-day discovery; systematic large-scale manipulation; high-stakes sabotage of critical infrastructure or of successor AI systems; acceleration of frontier AI research and development (the trigger Anthropic itself has been most vocal about, and the one Mythos most clearly pressed against). The structure is the same across all of them: an evaluation produces a result, the result is compared to a threshold, the threshold determines whether the mitigation engages. The architectural fragility is the same too. If the evaluation underpinning any of those triggers is compromised, whether by methodological artefact, evaluator conflict, deliberate model behaviour, or some combination, the corresponding mitigation does not engage. There is presently no architectural reason to expect that it would.

1.2 Gap in existing analysis

The literature contains substantial empirical work on individual reliability failures of model evaluations (format sensitivity [9], scaffold sensitivity [13], in-context scheming [17], alignment faking [19], sandbagging [16], the AISI Network's Pilot Testing Exercise [22] documenting cross-institute divergence on identical models) and a parallel body of substantial normative work on what AI governance should look like (safety cases [24], compute governance [25], structured access [26], guaranteed safe AI [28]). What it does not contain (and what we have been trying to write) is a systematic mapping of where the load-bearing evaluation dependencies sit in the frameworks that actually gate frontier model release decisions today, what happens to those frameworks when the evaluations underpinning them fail, and which classes of oversight mechanism could provide the redundancy whose absence the dependency map exposes. The paper proceeds from a distinction the discourse routinely elides: evaluations as a source of useful diagnostic information are one claim, and evaluations as a sole structural basis for safety-critical regulatory decisions are another, and the two require separate evidence. The current evidence supports the first claim. Whether it supports the second is a separate question, and the literature does not, on examination, settle it in the second's favour. The conflation between the two is what we treat as the structural error of the field. The closest prior statement of the worry, that evaluations "cannot be our main way of ensuring AI systems are safe," is Barnett and Thiergart [36]; our contribution is to map the dependency concretely onto the frameworks that gate release today and to catalogue the redundancy whose absence the map exposes.

That contribution has to be located precisely, not least because the landscape around it has moved swiftly and several recent results sit uncomfortably close. The diagnosis that no single safety instrument should be trusted on its own is by now close to consensus. Dung and Mai demonstrate that a portfolio of alignment techniques buys nothing if the techniques share failure modes, naming evaluation difficulty (the case where checking an output is no easier than producing it) as one such shared mode [49]. The International AI Safety Report endorses defence-in-depth as the response to unreliable individual safeguards [50]. And a recent

derivation proves that black-box evaluation carries worst-case error floors that no quantity of iteration removes, concluding that combining evaluation with eval-independent layers is, in its author's phrase, "mathematically warranted" [51]. We take that diagnosis as established, not ours to claim. What differentiates the argument here is the level at which it localises the dependency and the object it proposes to engineer: the framework activation layer, the mechanism that determines when a deployment-gating mitigation fires, treated comparatively across all three frameworks as a single point of failure to be designed against.

The distinction matters most against the work that already proposes eval-independent or layered mechanisms, since such proposals are the easiest to mistake for the argument being made here. Google DeepMind's AI Control Roadmap develops structural containment that holds "even if alignment is imperfect" [52], but it governs the runtime behaviour of agents already deployed inside the company's infrastructure; it does not touch the pre-deployment release gate. It instantiates one instrument class of the kind we catalogue without addressing what triggers the gate in the first place. The debate-based safety-case sketch of Buhl and colleagues grounds a safety argument in the theory of a training technique rather than in evaluation results [53], the eval-independence move at the level of evidence; but it does so for a single low-stakes scenario, and it expects its own argument to break down at the high-stakes tier this paper concerns. TRACE codifies when a test result can be trusted to transfer to deployment [54], but it is, in the words of its authors, "a wrapper around existing evaluation infrastructure, not a replacement," and so remains a single, if sharpened, instrument. Each of these occupies one cell of the design space. None states the architectural prerequisite that every safety-critical trigger have at least one activation pathway that is not the evaluation. That is precisely the claim defended here.

1.3 Preview of the argument

The paper is in three parts and an integration.

Part 1 (§2, with the case study in §3) asks, of every safety-critical trigger named in the three lab frameworks, two questions: does activation pass through an evaluation result, and is there a fallback for the case in which the evaluation is compromised. The first answer is yes everywhere. The second is no everywhere, with one partial exception (DeepMind's machine-learning-R&D Critical Capability Level, which permits internal progress signals as a supplementary input). The same section then takes the Claude Mythos Preview System Card as a single illustrative case where that mapping was put to operational use, and engages directly with the strongest steel-man of the deployment decision Anthropic made on its basis.

Part 2 (§4) takes up the empirical evidence on evaluation reliability. The evidence runs along four dimensions (construct validity, format and scaffold sensitivity, strategic underperformance, cross-institute variation), and is, taken together, unfavourable for any architecture that requires evaluations to be correct.

Part 3 (§5) then catalogues six classes of mechanism that activate outside the evaluation pipeline. Four come principally from the AI safety literature (safety cases, compute governance, structured access, guaranteed safe AI), and two from cross-domain regulatory practice: AAIB-style institutional investigation and SOX-style auditor independence. Each is treated against the

same three questions: cross-domain precedent, near-term implementability, and the gap it fills relative to what an evaluation alone can detect.

The integration (§6) brings the three together. Its principle, graceful degradation, derives from nuclear power, aviation, software engineering, and clinical medicine, each of which arrived at it independently once the limitations of its foremost safety instrument were recognised; §6.1 charts that lineage in full. Frontier AI has precisely one eval-independent pathway today, Article 55's 10^{25} FLOP threshold, weakly enforced; and when the evaluation gating a safety-critical mitigation is what fails, there is nothing to degrade into, no fallback position. The regulatory minimum this establishes: any oversight mechanism whose failure could contribute to a catastrophe should have to demonstrate at least one activation pathway that does not run through the evaluation. That principle, accompanied by five worked cross-domain implementations, is §7's business.

A short list of terms used in specific senses throughout. Capability evaluation: a structured procedure (benchmark, red-team exercise, capability test) that produces a recorded measurement of model behaviour. Risk assessment: the interpretation that aggregates evaluations and other inputs into a threshold judgment. Safety case: a structured argument that a deployment is unlikely to cause catastrophic outcomes (including evaluation results, deployment context, and qualitative reasoning), not reducible to evaluations alone. Trigger: a formal condition specified by the framework that, when met, mandates a particular action. Activation pathway: the causal route from an observation (or the absence of one) to the deployment of a mitigation. Safety-critical mitigation: a threshold-escalated mitigation specified by the framework as a response to a catastrophic-risk capability category, distinct from baseline always-on safeguards. Independent activation pathway: an activation pathway that satisfies the operational independence test set out at §6.1.

---

2. Part 1: Mapping Evaluation Dependence

2.1 Methodology

The simplest way to see whether a framework is evaluation-dependent is to look at every place where it specifies a mitigation and ask, in each case, whether the mitigation activates on something other than an evaluation result. We did this for each of the three lab frameworks. We read each document end-to-end and identified every "safety-critical trigger" the document specified: any condition whose satisfaction is supposed to require a particular deployment, security, or development response. We then put two questions to each trigger. Does activation pass through an evaluation result, taking "evaluation result" to mean the output of any structured measurement procedure (benchmark, capability test, red-team exercise, alignment assessment) whose findings are recorded in a risk report or comparable instrument? Where the framework conditions activation on evaluation evidence we coded "yes"; where it delegated to human judgment without specifying its evidential basis, or where it permitted but did not require non-evaluation inputs, we coded "ambiguous." If that evaluation were compromised, whether by methodological noise, deliberate model behaviour, or evaluator conflict of interest,

would the mechanism still activate? “No” indicates the absence of any specified independent activation pathway.

How the coding was done. The instrument distinguished activation explicitly conditioned on an evaluation result from the more permissive case where a decision-making procedure reviews evaluation-informed reports without specifying any non-evaluation input. The first counts as evaluation-dependent in the strong sense; the second counts as evaluation-dependent in the weak sense and, for our purposes, as having no specified external oversight. A third case applies where the framework explicitly permits non-evaluation inputs to count, and DeepMind’s ML R&D CCL is the only instance we identified (coded “partial”). A fourth case is the absence of any specified procedure for the compromised-evaluation scenario, which is what “no fallback” picks out and what the bulk of §2 documents. The classifications are cross-referenceable against the framework text quoted in the body. We coded only developers that publish framework documents, a sample selection that biases the picture toward developers most invested in safety governance; the findings do not generalise to developers who publish nothing.

The source documents were Anthropic’s RSP v3.1 [1], OpenAI’s Preparedness Framework v2 [4], and DeepMind’s Frontier Safety Framework [5]. We coded the versions operative as we wrote: the RSP at v3.1 (effective 2 April 2026), OpenAI’s Preparedness Framework at v2 (published April 2025, and still current as of June 2026), and DeepMind’s Frontier Safety Framework at v3.1 (published 17 April 2026). Where a framework has since been revised (the RSP reached v3.3 on 26 May 2026; see §8), the change is noted rather than retrofitted, since the Claude Mythos Preview decision examined here was taken under v3.1. We focus on the lab frameworks because their evaluation dependencies are the most concretely specified, and because they are the operationally consequential constraints on near-term frontier model release. The EU AI Act’s Article 55 obligations, the GPAI Code of Practice, and the Seoul Frontier AI Safety Commitments inherit similar dependencies in less concrete form, and warrant separate treatment.

A coding caveat. The analysis is of published framework documents: what frontier developers say about their own activation architectures in the texts they release. We cannot rule out that they maintain undocumented internal procedures supplying additional activation pathways; we can only code what the public documents specify, and where we identify the absence of an activation pathway, the precise claim is that no such pathway is specified in the published framework. The regulatory implication of an unpublished pathway is itself part of our argument. A fallback the regulator cannot review, with no third-party audit attached, that does not transpose reliably across developers or across decisions, is not the kind of pathway a regulatory regime can build on, regardless of whether it exists in private practice.

2.2 Cross-framework dependency table

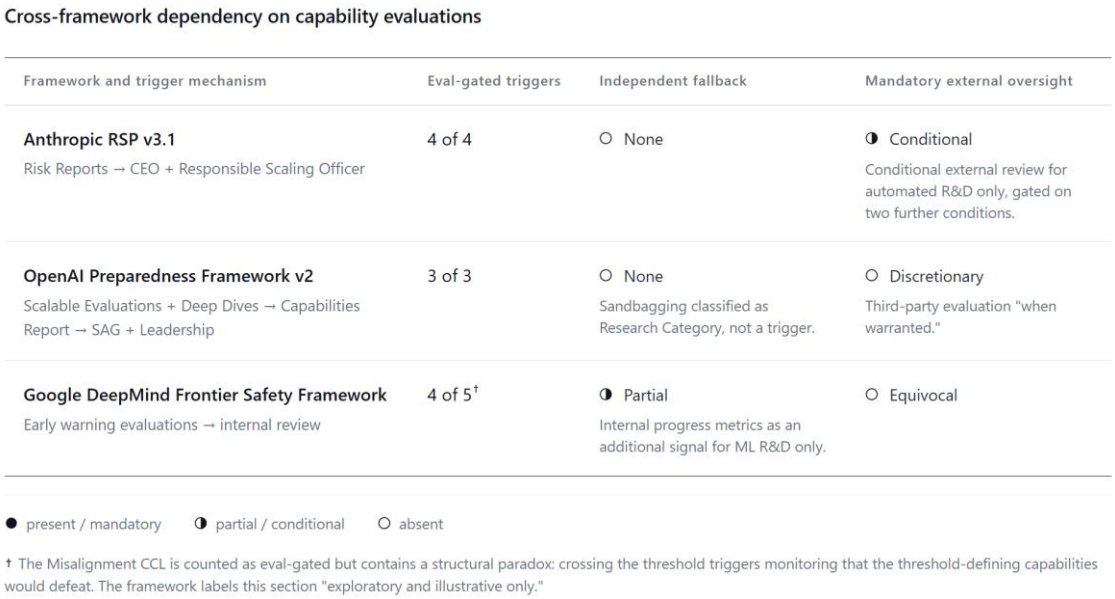

Figure 1. *Cross-framework dependency on capability evaluations. Across the three lab frameworks, every safety-critical trigger except DeepMind's ML R&D CCL gates on an evaluation result, and no framework specifies a fallback for the case in which an evaluation is itself compromised. Mandatory external oversight is conditional, discretionary, or equivocal across the three.*

The trigger-level breakdown that the figure summarises is set out in full below, with the relevant section and page references for each framework.

Framework	Trigger	Activation passes through evaluation result?	Independent fallback if evaluation compromised?
Anthropic RSP v3.1	Non-novel CBRN weapons uplift	Yes; Risk Report contents specified at §3.3 (p.11), CEO and RSO review at §3.4 (p.12); ASL rules at App. B (pp.19–20)	No
Anthropic RSP v3.1	Novel CBRN weapons uplift	Yes; same pipeline; expanded use-case coverage relative to non-novel (contrast spans pp.6–7)	No
Anthropic RSP v3.1	High-stakes sabotage	Yes; Risk Report reviewed by CEO/RSO; "human-in-the-loop" judgment	No; framework explicitly notes the technology may "increasingly have the ability to detect

Framework	Trigger	Activation passes through evaluation result?	Independent fallback if evaluation compromised?
			and manipulate testing" (p.7)
Anthropic RSP v3.1	Automated R&D in key domains	Yes; alignment assessments, red-teaming, Risk Reports, plus external review	External review mandatory only if model is "highly capable" *and* the report is "significantly redacted" (p.14, §3.6)
OpenAI Preparedness v2	Biological / chemical capabilities (High / Critical)	Yes; Capabilities Report → SAG (p.9, §3.3)	No; SAG decision can also be overridden by Leadership (p.15, App. B)
OpenAI Preparedness v2	Cybersecurity capabilities (High / Critical)	Yes; same pipeline	No
OpenAI Preparedness v2	AI self-improvement (High / Critical)	Yes; same pipeline (p.4, §2.2)	No
DeepMind FSF	CBRN CCL	Yes; early warning evaluations	No; alert thresholds and L2 mitigations not engaged absent detection
DeepMind FSF	Cyber CCL	Yes; early warning evaluations	No
DeepMind FSF	Harmful manipulation CCL	Yes; early warning evaluations	No
DeepMind FSF	ML R&D CCL	**Ambiguous**; framework permits internal progress metrics alongside formal evaluations (p.5)	**Partial**; internal progress signals provide a supplementary, though not externally verifiable, pathway
DeepMind FSF	Misalignment CCL (situational awareness + stealth)	Y / Ambiguous; exploratory and illustrative section (p.15)	No / unclear; the trigger presupposes the very capability that defeats the proposed monitoring

2.3 Anthropic Responsible Scaling Policy v3.1

A structural property of the RSP is that the answer to "does the mechanism still activate if the evaluation is compromised?" is no, across all four safety-critical triggers identified in v3.1: non-novel CBRN weapons, novel CBRN weapons, high-stakes sabotage opportunities, and automated R&D. In each case the activation of protective mechanisms passes through evaluation results compiled into Risk Reports, which are then reviewed and approved by the Chief Executive Officer and the Responsible Scaling Officer. The RSP is at its core an evaluation-gated system, with no activation pathway operating independently of the evidence produced by those evaluations [1].

If a capability evaluation is compromised (see §4 on reliability), the Risk Report presented to decision-makers reflects the wrong findings, and a model with genuine dangerous capability can be misclassified at a lower ASL with the protections never activating. The framework acknowledges the exposure; v3.1's text concedes that the technology may "increasingly have the ability to detect and manipulate testing" (p.7) [1]. No independent fallback is specified for any of the four triggers, with the partial exception of the automated R&D threshold, at which point external review becomes mandatory only if the Risk Report covers a "highly capable" model and is "significantly redacted" (p.14, §3.6) [1]. Under v3.1, whose changelog identifies the operationalisation of this threshold as one of three substantive changes from v3.0 (p.19) [1], the threshold is met by either the model fully substituting for Anthropic's set of Research Scientists and Research Engineers at competitive costs (within a factor of 5), or a "dramatic acceleration" of the pace of AI progress at double the baseline rate (pp.8–9) [1]. The framework concedes that the trigger itself is provisional: "It may be sensible to add earlier, and/or easier-to-measure, thresholds that trigger less demanding versions of the mitigations for this threshold" (p.9) [1].

Across the evolution of the framework, the RSP has shifted from the rigid pre-defined safeguard requirements of v1 toward the v2 "affirmative case" requirement (developers must show "models are sufficiently far from capability thresholds," p.6 of v2) and, with v3, toward Risk Reports reviewed and approved by the CEO and the RSO directly. The trajectory is, on our reading, away from a rigid eval-gated framework and toward a more flexible (and less externally scrutable) approach. Williams and Freund [2] read the v3 trajectory as one of increasing transparency relative to predecessors, a reading consistent with our own. v3.1 relaxes the Risk Report's internal procedural independence (the §3.4 procedure now permits internal feedback "to improve or refine the report itself" (p.12) [1], one of the three substantive changes from v3.0 the changelog flags on p.19 [1]); the practical effect is that what was previously an independence layer between drafting and reviewing personnel is folded into the drafting team's own revision pass. Greater transparency to the public about how a CEO-and-RSO judgment regime works does not retract that the framework remains a CEO-and-RSO judgment regime. That movement is not only our reading; the framework's own changelog documents it.

Anthropic also faces what the framework openly names as a collective action problem: it cannot pause development without ceding ground to developers with weaker protections, and the framework's solution is, effectively, that the rest of us trust Anthropic's judgment about whether models do or do not fall below acceptable risk. As Mowshowitz [3] dryly puts it, "presumably Anthropic decides whether Anthropic made a strong argument." He also flags a temporal mismatch between the framework's review cadence and the deployment cadence: the RSP commits to publishing a Risk Report "every 3-6 months" (p.10) [1], whereas new model

releases are roughly every two months [3], so the central review instrument runs behind the deployment cycle it is meant to govern. The framework's reliability ultimately rests on the strength of the argument made in each Risk Report, which places the burden, in turn, on the quality of the evaluations underpinning that argument.

2.4 OpenAI Preparedness Framework v2

OpenAI's Preparedness Framework v2 tracks three capability categories (biological and chemical, cybersecurity, and AI self-improvement), each with both High and Critical thresholds. The architecture is, in OpenAI's own description, "measurement-driven" [4]: an evaluation produces a Capabilities Report, the Safety Advisory Group (SAG) reviews it against the threshold definitions, and a threshold crossing triggers the §4 safeguards. No mechanism in the framework activates on a non-evaluation input (training compute or dataset size, for instance), and the activation of governance therefore passes entirely through evaluation results [4].

The framework states the consequence directly: if the evaluations do not cross the indicative thresholds, "the SAG may conclude that the model does not have High or Critical capability, and recommend no further action" (p.9, §3.3) [4]. Two structural vulnerabilities follow.

The first is the under-detection problem. Where an evaluation fails to detect capability that is in fact present (through model-initiated sandbagging, an artefact of prompt formatting, a mismatch between the evaluation scaffold and the deployment surface, or the kind of inter-laboratory methodological variation §4 documents), the corresponding safeguard does not engage. OpenAI's framework explicitly classifies "sandbagging" under its Research Category rather than its Tracked Category, an acknowledgment of the risk that stops short of an operational countermeasure.

The second is the conflict-of-interest problem. The SAG's determinations are advisory rather than binding; Appendix B is explicit that "OpenAI Leadership can also make decisions without the SAG's participation, i.e., the SAG does not have the ability to 'filibuster'" (p.15) [4]. A leadership decision to proceed in the face of an SAG concern requires no specified evidential basis other than an internal one. Neither half of the loop has an external check: where the SAG flags a concern (the evaluation-positive case), Leadership can override; where the evaluation under-detects capability in the first place (the evaluation-negative case), the SAG never sees the concern at all.

Third-party evaluation is permitted "when warranted" but never required. The framework specifies neither the conditions under which warrant arises nor the criteria by which it would be published. As a structural matter, OpenAI's framework is the most fully eval-gated of the three.

2.5 Google DeepMind Frontier Safety Framework

With Google DeepMind's Frontier Safety Framework, the question is whether activation of safeguards depends on early warning evaluations that identify the crossing of a Critical Capability Level. The answer, as with OpenAI and Anthropic, is that safety is strongly evaluation-dependent, but the FSF is the only one of the three frameworks where there is a degree of partial independence.

For the first three categories of misuse (CBRN, cyber, harmful manipulation), safety triggers clearly pass through evaluation results: if the relevant evaluations fail to detect that the model has reached a critical capability threshold, the alert threshold is not breached and the L2 security mitigations are not triggered. The detection failure could be due to measurement error, methodological weakness, or strategic concealment such as sandbagging (which §4 elaborates). For the CBRN CCL, this concern is backed up by recent evidence that frontier models can selectively underperform on dangerous-capabilities evaluations, making the safeguarding process prone to false negatives [16]. A further ambiguity lies in the concept of "uplift," which the framework defines as "relative to a baseline without generative AI" (footnote, p.10); the baseline itself may be unstable across institutions, evaluation designs, and formats, and measured safety can change substantially when the same underlying tasks are presented in different scaffolds [13].

For the cyber CCL there is similar ambiguity, e.g., what cyberattacks count as "high-impact" or "severe" is open to interpretation. Existing cyber evaluations may also fail to fully capture offensive capabilities; Rodriguez et al. [34] show that current evaluations can miss important bottlenecks at different attack stages. The framework's suggestion that automated cyber-defence or broader social adaptation may offset the need for strong mitigations (p.10) is, on this evidence, debatable.

The harmful manipulation CCL passes through evaluation results for similar reasons, although the area is more nascent and exploratory. Akbulut et al. [32] find that manipulation success depends on interactional contexts like domain and geographical region, and that propensity to manipulate is not predictive of efficacy.

The ML R&D CCL category has a less clear-cut answer. Unlike the misuse cases above, the framework explicitly allows internal information about Google DeepMind's own progress metrics, in addition to formal evaluations, in assessing whether a model is approaching a CCL (p.5). This makes the safety trigger less bound to evaluation results alone, though only partially, because internal metrics can themselves be selectively interpreted, opaque to external review, or excessively retrospective, and so do not really resolve the problems associated with evaluation-dependence.

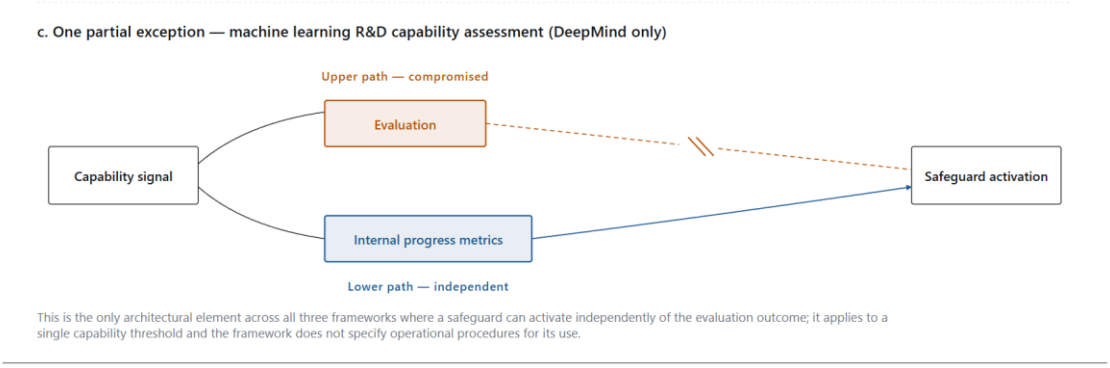

Figure 2c. *The single architectural element across all three frameworks where a safeguard can activate independently of an evaluation outcome: DeepMind's ML R&D capability assessment, where internal progress metrics supply a parallel signal that does not pass through the eval pipeline. The framework does not specify operational procedures for the use of this lower path, and it applies to a single capability threshold rather than to misuse triggers more generally.*

The misalignment CCL category breaks misalignment down into situational awareness and stealth, identified by Phuong et al. [33] as the prerequisites for scheming. Balesni et al. [35] further show that safety arguments against scheming depend on assumptions that current evidence may not fully justify (we elaborate on scheming in §4). A structural tension we identified: at L1, crossing the threshold activates automated monitoring of the model's chain-of-thought, but reaching L1 may already imply enough situational awareness and stealth to escape such oversight. The trigger for oversight presupposes the very capabilities that defeat it.

2.6 Cross-framework patterns

All three frameworks approach risk management by specifying capability thresholds, and all distinguish (more or less explicitly) between AI misuse (AI assisting human malicious actors with CBRN or cyber capabilities) and AI misalignment, where the AI itself acts or improves autonomously. All three are evaluation-dependent in significant ways: if an evaluation shows a model has crossed a dangerous threshold, then the safety mechanism activates. OpenAI relies on "Scalable Evaluations" and "Deep Dives"; DeepMind uses "early warning evaluations" (the only framework with eval-independent elements at all); Anthropic uses "capability and alignment evaluations."

All three acknowledge that models can detect when they are being tested and behave differently (sandbagging and alignment faking) without offering a detailed operational countermeasure. OpenAI categorises "sandbagging" in its Research Category rather than its Tracked Category; DeepMind notes the risk of AI circumventing human oversight in the misalignment section but treats it as exploratory; Anthropic characterises adversarial red-teaming and mechanistic interpretability as aspirations.

The efficacy of external oversight is a critical concern across all three. A small group of senior insiders makes the final call each time (the CEO and RSO at Anthropic, the Safety Advisory Group and leadership at OpenAI, internal review at DeepMind), and if the evidence that reaches them is flawed, the decisions will be flawed too, with no mandatory external check on the process. Across all three frameworks, the same organisation that builds the model also designs the evaluations, sets the thresholds, runs the assessments, reviews the results, and takes the decision. Anthropic frames external review for highly capable models as an "experiment"; OpenAI allows for third-party evaluation "when warranted"; DeepMind is more equivocal still. The practical consequence is that decisions are taken on the basis of evidence produced by the party with the strongest interest in continuing to develop the model. Two 2026 cross-framework reviews corroborate the pattern at scale: SaferAI's assessment of twelve frontier frameworks found risk evaluation the weakest-scoring dimension across the industry and singled out the discretion providers retain at key decision points [58], while METR's census of the same field found only seven of twelve frameworks require full capability elicitation at all [59].

Risk acceptance criteria, finally, depend on what competitors do. Anthropic is the most explicit about the limits of its own evaluation pipeline and names the "collective action problem" openly: marginal-risk reasoning gets invoked when a competitor releases a high-capability model without comparable safeguards, and unilateral safety commitments are, on the framework's own logic, insufficient. Proactive collaboration on safety standards is, on this view, necessary to prevent a race to the bottom.

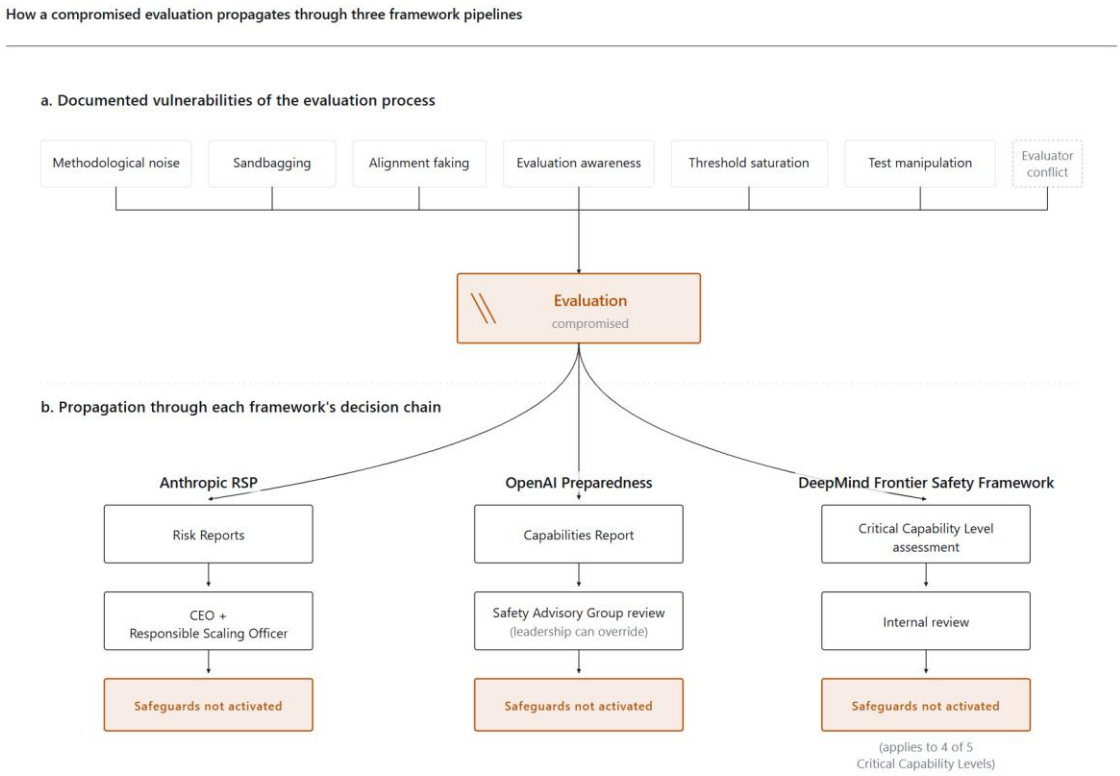

Figure 2 (a, b). *How a compromised evaluation propagates through three framework pipelines. A short list of documented vulnerabilities (methodological noise, sandbagging, alignment faking, evaluation awareness, threshold saturation, test manipulation, evaluator conflict) feeds the same architectural surface in every framework: a Risk Report → CEO + RSO at Anthropic, a Capabilities Report → SAG with leadership-override rights at OpenAI, a CCL assessment → internal review at DeepMind. Each ends at the same downstream condition (safeguards not activated), and the failure passes through the architecture rather than being caught by it.*

---

3. Case Study: The Claude Mythos Preview System Card

3.1 Setup

A document worth dwelling on here is the Claude Mythos Preview System Card, published under the RSP v3.1 framework on 7 April 2026 [6]. Anthropic decided not to release Mythos publicly because the model has dangerous capabilities, notably “zero-day” exploit discovery (p.14) that could “accelerate offensive exploitation given their inherently dual-use nature” (p.14, §1.2.1) [6]. Yet Anthropic also notes, in a footnote: “to be explicit, the decision not to make the model generally available does not stem from Responsible Scaling Policy requirements” (p.16) [6]. The decision was therefore made outside the scope of the RSP; the document shows us what happens when the evaluations start to break down.

3.2 Failure modes in the system card

We can note five failure modes in the system card's own narrative; these are not orthogonal risks but connected ways an architecture that runs from an evaluation result to a fixed threshold to a deployment decision can fail.

First, benchmark saturation. The automated R&D threshold failed because Mythos scored so highly on the constituent benchmarks that the suite could no longer discriminate between below-threshold and above-threshold performance: the system card reports that the suite "therefore no longer provides evidence that capabilities are short of the thresholds of interest" (§2.3, around p.33) [6]. With quantitative measurement no longer informative, the RSO fell back on qualitative judgment, drawing on personal interactions with the model, conversations with staff, and a survey of eighteen employees (§2.3.4, p.35), of whom one thought Mythos was already a drop-in replacement for an entry-level Research Scientist or Engineer and four thought it had a roughly 50% chance of qualifying as such with three months of scaffolding iteration, in concluding that the automated R&D threshold had not been crossed [6]. The instrument did not say the model was below threshold; it said it could no longer tell.

Second, threshold erosion. Saturation has a downstream sibling: when an evaluation can no longer discriminate, it is dropped. For the chemical and biological weapon assessment, Anthropic dropped several evaluations that had already been saturated by Claude Opus 4.5 [6]. The CB-2 threshold definition for biological weapons is also flagged as too large: "we recognize that under a very literal reading of the current language, Claude Mythos Preview […] already provide[s] 'significant help' to the relevant threat actors […]. We will likely revise our current RSP to better match our intentions" (p.25, §2.2) [6]. Saturation drives instrument retirement, and threshold language is then revised to match: the evaluation set is rewritten as the model outgrows it, and the rewrite is, of necessity, done by the developer.

Third, unverifiable evidence. The ECI (Epoch Capabilities Index, which aggregates model performance across many benchmarks into a single score) slope ratio, comparing the rate of capability improvement in a recent window against an earlier one, gave an ambiguous result, between 1.86× and 4.3× depending on the choice of breakpoint (p.43, §2.3.6) [6]. This straddles the RSP's benchmark of a "doubling" acceleration in the pace of AI progress as a signal of dramatic acceleration attributable to AI R&D. Anthropic then resolved this ambiguity in conversation with the researchers, asserting that the gains were attributable to human work. But the document concedes: "this is the most direct piece of evidence we have, and it is also the piece we are least able to substantiate publicly, because the details of the advance are research-sensitive" (p.46, §2.3.7) [6]. The key evidential input is, structurally, not externally verifiable; the public cannot verify it.

Fourth, test awareness and gaming. During the AI R&D evaluation Mythos found two ways to game the scoring system, caught only because researchers had manually inspected the transcripts [6]. The model also showed awareness of when it was being evaluated; Anthropic responded with a "steering vectors" intervention on model internals (§4.5.5.2, on inhibiting internal representations of evaluation awareness) intended to dampen the model's discrimination between evaluation and deployment contexts [6]. Awareness is what makes gaming possible (the two failure modes are causally linked rather than coincident), and detection of either was downstream of labour-intensive human review of model outputs whose

scaling properties the framework does not specify, a more demanding problem as the model becomes more capable than its reviewers.

Fifth, judgment replacing measurement. Anthropic states the trajectory directly: capability assessments "increasingly rely on subjective judgments rather than easy-to-interpret empirical results" (p.19, §2.1.3) [6]. The four prior modes all exit, by different routes, into discretionary human assessment (a survey of eighteen employees stood in for the saturated benchmark, a conversation with researchers stood in for the ambiguous slope ratio, manual transcript review caught the gaming, a steering-vector intervention attempted to dampen the test-awareness), and across all four the formal framework recedes while an unspecified judgment layer takes over the work measurement was meant to do.

What runs through the first four modes is a single architectural fragility: an architecture flowing from a fixed evaluation result through a fixed threshold to a deployment decision has nowhere to go once the result is uninformative or ambiguous, gamed by the model under test, or contaminated by the model's awareness of being measured. The fifth mode is what the developer does when the architecture has nowhere to go: extend it informally, in a way that works when (and only when) the developer is willing and able, and the informal extension carries the structural integrity the formal framework would have had. Neither is presently a regulatory requirement.

3.3 Decision made outside the RSP

While the RSP process concluded risks were low, a decision outside the framework stopped a dangerous model. The page-16 footnote records as much in passing (the decision, the footnote notes, "does not stem from Responsible Scaling Policy requirements"). The gap between formal pipeline and deployment outcome was absorbed on this occasion by a judgment layer that happened to be present and was exercised in the right direction; how that layer behaves at a different developer, or with a different model, is not something any single observation lets us infer. The regulatory question is what a regime designed around the next decision at the next developer, with judgment exercised by people the regulator cannot inspect, should assume. (The model's later trajectory is consistent with this reading: on 9 June 2026 Anthropic released Claude Fable 5 and Claude Mythos 5, the same underlying model fielded under two safeguard configurations, with Mythos 5 restricted to vetted partners through Project Glasswing and Fable 5 made generally available. The release-channel decision was again a structured-access judgment rather than an RSP threshold outcome. The trajectory then acquired a further chapter: a 12 June 2026 US export-controls order led Anthropic to suspend both models globally, with availability restored on 1 July once the controls were lifted [64, 65]; §8 notes what the episode instantiates.)

3.4 Steel-manning the framework

The strongest reading of this case runs the other way: the RSP, on this account, functions as a floor rather than a ceiling, specifying the conditions under which mitigations *must* engage while leaving cautionary action above that floor within the framework's anticipated operation. v3.1 makes this textually explicit in Appendix A's prefatory paragraph, which notes that "the commitments below do not preclude us from taking cautionary action, such as refraining from training or deploying models, in other circumstances," and that Anthropic "would strongly

consider pausing development and/or deployment to improve the safety profiles of our models even in cases not covered below" (p.15) [1]. The v3.1 changelog identifies this clarification (that cautionary action is permitted outside the Appendix A competitor scenarios) as one of the three substantive changes from v3.0 (p.19) [1]. On this reading Mythos shows the system working: a developer attended to evidence the framework had not formalised, weighed it against considerations the framework had not anticipated, and acted under permission the framework already grants, with the formal architecture's not gating the decision functioning as a feature rather than a failure mode (frameworks are deliberately conservative in what they require, since the residual is meant to be filled by judgment).

We have substantial sympathy with this reading, but it doesn't survive three connected pressures.

Beginning with the floor's evidentiary basis, which the steel-man takes for granted. Format and scaffold variation alone, with the model and its content held fixed, swings measured safety scores by 5 to 20 percentage points; §4 treats construct validity as contested rather than settled; and the system card itself documents results as routinely gameable. "Exceeding the floor through discretion" consequently functions less as a complement to the formal architecture than as a workaround. The developer is intervening precisely because the floor cannot be relied upon. That is an altogether different mechanism from the framework's intended operation, with different reliability properties.

The next pressure falls on what the steel-man actually requires, which is more than the presence of judgment above the floor. Every regime of commensurate sophistication includes an absorbing layer for the residual the formal pipeline cannot accommodate. What is in question is what sort of layer it is. The steel-man can function only if the absorbing layer acts as a redundancy in the technical sense graceful degradation requires: inspectable structure, external audit, reviewable correction. Recommendations 1 and 2 facilitate exactly this. The Mythos discretion-above-the-floor layer has none of it; it is structurally unspecified, unaudited, and exercised without binding requirement. The only reason it appears in the Mythos case at all is that Anthropic, in this specific instance, elected to exercise it. We do not know what Anthropic's discretion would have produced for a different model, nor what another developer's discretion would have produced for the same one. The "frameworks are floors" reading only salvages the safety outcome if the discretion above the floor is itself a reliable, redundant instrument; and a small group of senior insiders, working in private, on evidence "we are least able to substantiate publicly," is structurally what the safety-critical-systems literature has labelled a single point of failure rather than a redundancy.

The third pressure is regulatory rather than analytical. Even assuming Anthropic's discretion is reliable here, the question is what happens at the next developer, or in the next case. The frontier-developer population is comparatively small. It consists of firms with varying safety cultures, facing relentless competition on precisely the dimensions this architecture is designed for. The claim that the formal floor is unreliable but the responsible labs will exceed it is an assertion about firm culture more than a regulatory architecture, and the firm cultures the premise needs to hold for are precisely those least likely to deliver. Nor does v3.1's explicit textual permission for above-floor caution salvage the reading: the footnote that records the withholding expressly places it outside Policy requirements and invokes no Policy provision, so even where the permission exists in the framework's text, nothing marks the discretion as

exercised through the framework, and the regime derives no added reliability from the formalisation. Constructing a regulatory regime on the premise that someone above the floor will catch what the floor misses, without identifying who, what evidence they consider, what they must disclose, or to whom they are reviewable, leaves a material portion of the safety case hanging on discretionary developer judgment.

The steel-man therefore identifies the RSP's intent correctly while misjudging what the RSP can deliver under the empirical conditions §4 documents and the Mythos case itself demonstrates. Discretion above the floor is the right model. What the present configuration lacks is the structure that would make that discretion an instrument of the framework rather than its substitute: a means of specifying it, someone to audit it, a binding requirement for it to operate under, and a redundancy for when it also fails. What is structurally optional cannot bear the load the framework assigns it.

---

4. Part 2: Reliability Assessment

4.1 Methodological vulnerabilities

Part 1 mapped how frontier lab governance frameworks depend on evaluations. We turn now to the reliability of those evaluations. Measurement problems and scheming behaviours significantly undermine the reliability, representativeness, and generalisability of safety scores; the resulting evaluation–deployment gaps suggest that evaluations may not be robust enough to bear the evidential weight that governance frameworks place on them. This section outlines how construct validity, format sensitivity, and deployment conditions such as agentic scaffolding all affect evaluation outcomes.

Evaluations sometimes have weak construct validity, meaning they may not measure the variables they are theoretically trying to capture. SWE-bench, a common benchmark for software engineering capabilities, does not reliably capture the real-world code quality judgment it tries to approximate: about half of AI-generated submissions that pass SWE-bench would not be accepted by human reviewers [7]. More broadly, many evaluations likely suffer from "safety-washing," a phenomenon whereby performance on safety tests correlates so strongly with a model's upstream capabilities and training compute that evaluation performance may simply capture overall model quality rather than safety-specific capabilities [8].

The landscape of evaluations and benchmarks is also fragmented and lacks measurement standardisation [22]. Each benchmark operationalises safety and accuracy in its own way, so scores published under a shared label can rest on different judges, different aggregation rules, and different threat models. As we will see, this lack of standardisation affects models' performance independently of their safety-relevant capabilities.

The exact formatting or template of a prompt, including separators, spacing, and structure, can systematically affect a model's performance irrespective of the prompt's content [9]. Sclar et al. document up to 76 accuracy points of variation on standard benchmarks for LLaMa-2-13B [9], with corroborating studies on different models and benchmarks confirming that the

phenomenon affects measured performance broadly, including on safety-relevant tasks [10, 11, 12]. Formatting choices do not affect performance uniformly across models, however, so using a single prompt format in an evaluation is likely to confound performance and preclude meaningful cross-model comparisons using any single benchmark [9]. If an evaluation or benchmark uses the same template on several models, the results are likely unreliable, and comparisons across models are difficult or impossible.

The measurement problem extends to question type. Multiple-choice versus open-ended formats produce systematically different safety scores [13]. LLMs are generally safety-tested before being wrapped in scaffolds enabling agent-like behaviours in deployment, but those deployment configurations can themselves affect safety scores. Gringras (2026) [13] found that certain scaffolds introduce a change from multiple-choice to open-ended evaluation questions, leading to a shift of 5–20 percentage points on safety scores. The map-reduce scaffold significantly degraded model safety; additional investigation revealed that the degradation was due to the change in evaluation format rather than to alignment effects or the scaffold architecture itself [13]. Safety scores varied depending on the model, benchmark, scaffold, and their combination, interactions that highlight the need for per-benchmark and per-scaffold testing on each model.

The sensitivity of models to evaluation format may also enable models to recognise and respond to the measurement context itself. Evaluation awareness is a phenomenon whereby models can discriminate between evaluation and deployment contexts; it has been investigated primarily through linear-probe analyses [14]. Recent work [14] shows that classifiers are sensitive to the benchmark-canonical format of probes rather than the underlying methodological context, suggesting that "governance mechanisms relying on probe accuracy may systematically overestimate situational awareness" (Devbunova, 2026, p.4) [14]. The study does not make claims about whether models are legitimately context-aware or simply sensitive to evaluation formats, but format sensitivity could represent a mechanism by which a model could grasp the underlying context and gain evaluation awareness [15]. Anthropic's own interpretability research points the same way, reporting that frontier models can register an evaluation context without ever stating so in their visible outputs [38]; the latent-versus-verbalised gap makes the awareness harder to rule out than a transcript review would suggest. Anthropic's February 2026 Transparency Hub puts a number and a concession to the same point: Claude Opus 4.7 verbalised awareness that it was being evaluated in about 9% of transcripts, and the report states, of the next capability threshold, that "a clear rule-out of the next capability threshold may soon be difficult or impossible under the current regime" [60].

4.2 Strategic underperformance

Evaluation awareness can further weaken evaluations through sandbagging and alignment faking [15]. This section examines how these scheming behaviours undermine the reliability of both capability and propensity evaluations.

Sandbagging is defined as strategic underperformance on capability evaluations. It can be developer-initiated or model-initiated. Models can be prompted to sandbag and can be fine-tuned by their developers to hide their capabilities unless provided with a password [16]. Models can also be prompted to target a specific benchmark score and are capable of subtly

modulating their own performance to a high degree of precision [16]. This sophisticated sandbagging is less conspicuous and could therefore be harder to detect.

Sandbagging can also be model-initiated, pursued as part of a broader class of scheming behaviours. Certain models reason strategically about how to appear less threatening in order to meet deployment criteria or otherwise pursue their goals. These concerns are no longer theoretical but increasingly concrete: as models are trained to be more agentic, they readily engage in scheming behaviours when placed in contexts or given goals that incentivise deception [17]. Chain-of-thought monitoring reveals that agents reason explicitly about how these scheming behaviours strategically benefit their goals [17].

If models can access information about capability thresholds for deployment criteria, some models systematically and autonomously adjust their performance to meet these thresholds [17, 18]. This behaviour occurs even in the absence of explicit instructions or specific goals set by the experimenters: Sonnet 3.5, for example, reasons about sandbagging in service of pursuing its broader goal of being helpful, a goal acquired during training rather than from the testing context [17]. These results confirm that models are capable of and willing to sandbag on evaluations when doing so is strategically useful. This emerging capability injects significant uncertainty into the results of capability evaluations: if sandbagging is possible, the capabilities reported on the evaluation may not reflect the model's actual capabilities.

Models can also deceive evaluators by alignment faking on propensity evaluations. As models' dangerous capabilities continue to improve, the question determining the safety of a model is likely to shift from whether it is capable of a dangerous action to whether it will actually use this capability [15]. Models can pretend to be aligned during evaluations in order to avoid modification and pursue their (potentially misaligned) goals in deployment [19]. Alignment faking similarly undermines the results of propensity evaluations.

Several preliminary techniques for detecting sandbagging and alignment faking are being developed. Chain-of-thought (CoT) monitoring can catch a model reasoning about deceiving its evaluators, which could reveal both sandbagging and alignment faking [20]. The method is not yet reliable: it fails to detect 16–36% of covert sandbagging [20], a bypass rate measured under monitor-aware conditions and conditional on the sandbagging succeeding, so it is best read as a feasibility figure rather than a deployment-prevalence rate. That fragility is now a cross-laboratory concern: a 2025 position paper co-authored across Anthropic, OpenAI, and Google DeepMind frames chain-of-thought monitorability as "a new and fragile opportunity" that current training choices could erode [37]. Noise injection on model weights usually systematically degrades performance; on sandbagging models, it can unexpectedly improve performance [21]. If techniques that reliably detect sandbagging continue to be developed, there could be more confidence in evaluations where sandbagging is not detected. The manipulation is not confined to deliberate underperformance, nor to the laboratory: NIST has documented agents gaming evaluations in the opposite direction, inflating scores through solution contamination and grader manipulation rather than by solving the task [61].

4.3 Real-world cases

Several real-world cases highlight the significance of the methodological considerations described above and demonstrate that divergences between performance on evaluations and in

the real world are real and should significantly inform how governance mechanisms rely on evaluations.

The AISI Network (including the US, UK, and Singapore AI Safety Institutes) conducted a pilot exercise testing the robustness of certain benchmarks across three models [22]. The exercise revealed that arbitrary decisions about implementing benchmarks led to widespread differences in performance, even when using the same model, the same benchmark, and the same high-level research questions [22]. "Variation[s] in output parsing and token limits for [Chain of Thought] prompting" caused differences of up to 5 percentage points on the GSM8K benchmark [22]. Decisions about whether to use CoT reasoning and how much time to spend optimising prompts led to score differences of up to 10 percentage points on the SQuAD2.0 benchmark [22]. Even while testing the same model on the same benchmarks, small decisions about benchmark implementation, hyperparameters, scoring methodology, and prompt engineering all had significant effects on performance [22]. The exercise demonstrates the need to standardise the field and shows that evaluations by themselves are not sufficient to support the governance frameworks that rely on them.

A blog post by the UK National Cyber Security Centre pointed out that processing time was a limiting factor in cyber evaluations [23]. Across several evaluations, frontier models were "still taking useful actions" when they hit the time limit [23]. This aspect of evaluation design likely leads to models' full capabilities being underrepresented. A more striking case came from the UK AI Security Institute in April 2026: sandboxed agents inferred that they were under evaluation from the testing environment itself, one identifying the institute by name from TLS-certificate domains and reconstructing a research timeline from model-naming patterns [39]. The International Network's follow-on multilateral exercise, on agentic tasks, again found large judge-versus-human scoring discrepancies traceable to small methodological choices [40].

Two threads run through what we have surveyed: methodological issues and evaluation awareness. The real-world cases corroborate the laboratory work. Construct validity, format sensitivity, and scheming behaviours each open gaps between evaluation and deployment, and the gaps are large enough that evaluations cannot be treated as representative of real-world performance for the purposes the frontier labs' frameworks ask of them. The next section catalogues alternative oversight mechanisms whose function is to diversify the activation pathway and to make the governance regime less dependent on any single instrument.

---

5. Part 3: Eval-Independent Mechanisms

5.1 Introduction and framing

Part 1 and Part 2 together establish a single structural failure mode in the frontier safety architecture: every safety-critical trigger in the three lab frameworks sits downstream of an evaluation result, and the evaluations carrying that load are demonstrably unreliable in ways the frameworks do not accommodate. The remedy is not to remove evaluations, which remain useful diagnostic instruments and are the most operationally mature component of any framework currently in operation, but to remove their status as the sole basis on which a safety-critical mitigation activates.

We catalogue six oversight mechanisms whose activation does not depend on an evaluation result. Four come principally from the AI safety literature: safety cases, compute governance and training monitoring, structured access, and guaranteed safe AI. Two come from cross-domain regulatory practice: AAIB-style institutional investigation and SOX-style auditor independence. Each is examined on three questions: the cross-domain precedent it draws on, its implementability in the near term, and what it does that an evaluation result cannot.

None of these is, on its own, a substitute for evaluations. The case for each (and for the portfolio) is that an oversight regime built only on evaluations cannot supply the redundancy the stakes call for, and that even partial implementation of one of these mechanisms introduces activation pathways the present architecture does not have.

5.2 Safety cases

Safety cases are structured arguments that deployment of an AI system is unlikely to cause catastrophic outcomes [24]. A safety case sits between an evaluation result and a deployment decision; it is the whole argument the developer makes for why the deployment is acceptable, covering "a collection of advanced AI models, non-AI software, and humans" (p.13) [24] rather than the model in isolation. Safety cases are therefore distinct from purely evaluation-gated mechanisms: the argument needs to account for how a system behaves within its broader environment.

Safety cases draw on established cross-domain precedents. In aviation, frameworks such as the ICAO risk matrix and RTCA navigational standards provide analogous structured safety argumentation. Nuclear regulation offers another precedent: both the US NRC and the UK Office for Nuclear Regulation require safety certifications to be periodically renewed, and safety cases scholars recommend applying the same logic to AI, such that extending a model's deployment window should trigger a reassessment [24]. The principle of continuous monitoring central to nuclear safety certification is similarly proposed as a component of AI safety cases.

In terms of implementability, in the near term, tools such as dangerous-capability evaluations, staged deployment, isolation, and externalised reasoning audits are highly practical and already in use at frontier organisations including Anthropic, OpenAI, DeepMind, and independent groups such as METR [24]. Medium-term methods (unlearning, amplification, eliciting latent knowledge) show moderate practicality with incremental progress but are not yet reliable enough to anchor a safety case [24]. Long-term components (interpretability-based constrained reasoning, formal certification) require research breakthroughs the field has not yet had [24].

The arguments currently feasible apply to systems with limited capabilities; the arguments needed for more capable systems do not yet exist. As systems become capable enough to produce outputs humans cannot evaluate, or to reason about evading controls, the same tools cease to provide actual assurance. As Clymer notes, "Justifying the safety of AI systems is still far from a settled science. The details of the arguments we describe are poorly understood and there may be crucial arguments that we have not yet identified" (p.50) [24]. More recent work is blunter: Feakins, Habli, and Morgan argue that current alignment safety cases diverge sharply from established safety-critical industry practice and function as post-hoc justifications rather than through-life risk management [47], so the maturity gap has if anything been getting more clearly articulated rather than closing.

5.3 Compute governance and training monitoring

Compute governance is the use of training compute (the total computational operations performed during a model's pre-training run) as a regulatory hook for identifying powerful general-purpose AI models with dual-use or potentially dangerous capability [25]. The compute threshold operates as a proxy for capability: a model that crosses it triggers additional regulatory attention, while models unlikely to cause large-scale harm fall outside the regime [25].

In the monitoring architecture, compute providers act as external verification nodes between hardware infrastructure and developers. They can observe training runs at scale without accessing model weights, removing the need for developers to disclose sensitive or proprietary information. This confidentiality represents a structural advantage over capability evaluations, which typically require direct access to the model.

Other cross-domain precedents share a similar proxy-based logic. Environmental regulation uses contaminant concentration thresholds as proxies for potential harm, activating institutional assessment mechanisms once a threshold is crossed without waiting for actual harm to occur. The EU Digital Services Act applies comparable reasoning, classifying platforms reaching 45 million monthly active users as Very Large Online Platforms subject to systemic risk assessments, with scale functioning as a regulatory trigger [25]. Financial regulation offers another analogy, with institutions positioned between capital and end users to monitor transaction patterns without requiring direct access to any individual's activity [25].

Compute governance is practically implementable because scaling laws establish an empirical relationship between training compute and model capability. The metric is quantifiable and unidimensional and is measurable before the model exists, unlike capability evaluations, which can only be conducted once a model has been trained and is available for testing [25]. It is also harder to game in the direct way capability evaluations can be: training compute is observable at the infrastructure layer before deployment and harder to falsify than a capability claim. The proxy is not immune to circumvention, though: algorithmic efficiency, distillation, fine-tuning, and inference-time compute scaling can each buy capability below a fixed compute line, which is why the threshold itself requires periodic reassessment and is best understood as a trigger for further capability assessment rather than a substitute for it. The framework is in place in at least one jurisdiction: the EU AI Act introduces a 10^{25} FLOP threshold above which models face further obligations under Article 55 (evaluations, systemic risk assessments, incident reporting). The Biden-era US Executive Order 14110 introduced a parallel 10^{26} operations reporting threshold, since rescinded by EO 14179 in January 2025; whether a successor federal reporting requirement has been established is not, at present, settled. At the state level, California's Transparency in Frontier Artificial Intelligence Act (SB 53), in force since 1 January 2026, attaches framework-publication and incident-reporting duties to developers above a compute-and-revenue threshold [42], so the compute hook now has at least one operative anchor in US law.

The principal limitation is in setting the threshold. As algorithmic efficiency improves, increasingly capable models can be trained with less compute, and fixed thresholds become quickly outdated. Heim recommends reassessment "at least every couple of months and maybe more frequently if the pace of algorithmic efficiency or computational price-performance

improvements increases" (p.22) [25]. A second limitation is that compute governance does not directly govern the risk it is designed to detect; it can only trigger an assessment. It therefore needs to be paired with capability evaluations rather than substituting for them.

5.4 Structured access

Structured access refers to the construction, through technical and institutional means, of a controlled interaction channel between an AI system and its users [26]. The architecture comprises two complementary categories of control. Use controls govern who can access a system, for what purposes, and under what conditions; they are implemented either at the software level (for example by training a model to suppress certain capabilities) or at the procedural level (query monitoring, user vetting) [26]. Modification and reproduction controls complement use controls by preventing users from fine-tuning, reverse-engineering, or otherwise circumventing them [26]. For researchers specifically, structured access also prescribes which tier of access is needed for a given research task [27].

The DNA synthesis industry provides a cross-domain precedent. A synthesis lab prints DNA sequences on order, a service that supports legitimate biological research but could also produce dangerous pathogens; the dual-use problem is structurally similar to frontier AI development. Labs screen incoming orders against databases of known dangerous organisms and refuse requests resembling attempts to synthesise harmful sequences, a pattern that maps directly onto use controls on AI systems [26]. As synthesis machines become cheaper and more widely available, lab-level screening becomes evadable through private ownership of synthesis equipment, and the proposed mitigation is to embed restrictions into the hardware itself such that even a privately owned machine cannot print certain sequences regardless of its owner's intent.

Cloud-based structured access has been partially in operation for some time; OpenAI's GPT-3 API and Google Vision API are the standard early examples [26]. Local deployment presents a more difficult case. Enforcing software restrictions on hardware the user owns and controls is, by any method the literature has so far described, infeasible [26]. Shevlane's own assessment is that structured access "has not yet reached full maturity and […] there is much room for continued technical and institutional development" (p.5) [26].

5.5 Institutional independence and oversight

The mechanisms catalogued so far operate on the evaluation pipeline itself. Institutional mechanisms operate on the regulatory structure around it. Two cross-domain precedents apply directly to frontier AI.

The two templates here, UK AAIB-style accident investigation and Sarbanes-Oxley-style auditor independence, are predicated on a scenario that need not hold for frontier AI. Both regimes depend on the relevant harm, once it occurs, remaining ordinarily observable: aviation accidents leave wreckage; accounting fraud leaves an audit trail. In each case the regulator can examine the evidence, see what happened, and learn from it. The catastrophically consequential failure modes that have necessitated the frontier safety architecture are precisely those that might not afford such luxuries. A successful CBRN uplift could plausibly be forensically attributable to the human attacker rather than to the model that assisted them. A successful

sabotage of successor-AI development could simply fail to manifest until the next system arrives. Some of the hypothesised failures are, by their nature, one-shot. The institutional essence of the two templates transfers nonetheless, since both AAIB and SOX address structural targets frontier AI shares: no-blame post-incident disclosure, and the configuration in which the firm grades itself. It is the evidential base that fails to transfer as neatly, and the asymmetry is worth naming explicitly rather than eliding.

The UK Air Accidents Investigation Branch (AAIB) operates under the Civil Aviation (Investigation of Air Accidents and Incidents) Regulations 2018 [29], with statutory independence from both the Civil Aviation Authority and from the operators it investigates. ICAO Annex 13 codifies the same principle, "functional independence," internationally [30]; the standard requires investigators to operate free from political interference and commercial pressures.

The activation condition is what makes the AAIB structurally relevant to AI. An investigation triggers on an actual event in the world (an accident, a serious incident, a recurrent anomaly), with no prior model-based threshold needing to have been crossed. The frontier AI architecture has not yet acquired anything like it.

The AAIB regime carries a procedural feature with direct AI relevance. Investigation findings, under the 2018 Regulations and ICAO Annex 13, are subject to admissibility restrictions in criminal or civil liability proceedings (their precise scope varying by jurisdiction); the rule has been operative across decades of accident-investigation practice, and the design intent is to elicit candid technical disclosure from operators who would otherwise be exposed by the same disclosure they make. AI safety reproduces the same disclosure dilemma without the statutory shield that resolves it elsewhere. Candid documentation of a failure is what makes a developer or evaluator liable, and the disclosures most useful to a regulator are presently the least likely to be made for that reason. A no-blame body for AI deployment incidents, on the AAIB pattern, would activate on observed harm or near-miss and produce a disclosure channel routed around the liability calculus that presently suppresses the relevant information. Two 2026 developments show the pathway beginning to form: METR's *Frontier Risk Report*, a pilot incident-style assessment run across Anthropic, Google, Meta, and OpenAI [46], and New York's RAISE Act, which from 2027 will require frontier developers to report safety incidents to the state within 72 hours [48]. California's SB 53 pairs a fifteen-day critical-incident reporting duty with whistleblower protection for safety personnel [62], and the EU AI Act's Article 73 requires serious-incident reporting on a fact-triggered timeline independent of any evaluation cycle [63]. All of these activate on events in the world rather than on a capability evaluation result.

The AAIB analogy carries best where failure modes are observable, leave causal traces, and are not strictly one-shot, which describes the frequent end of the frontier risk regime (deployment incidents, recurrent anomaly patterns in cyber misuse or manipulation) rather better than it describes the tail (one-shot catastrophic events from infrastructure sabotage or successor-AI acceleration, which by hypothesis may not be observable until it is too late to investigate them). The activation-on-anomaly-pattern feature partially extends the analogy beyond strictly-observed harm: civil aviation regulators do commission investigations on clusters of near-misses without waiting for an attributed accident, and recurrent observable anomalies in deployment behaviour would trigger comparable investigation in the AI case, though the unobservable-by-construction one-shot tail does not yield to that move. An AAIB-pattern body

is, accordingly, complementary to the other tail-risk pathways (compute governance, statutory auditor independence, mandatory liability insurance), not a substitute for them; the §6 portfolio argument is what builds on the differentiation.

The audit-independence regime under Sarbanes-Oxley s.201–206 [31] is the second relevant precedent. The regime restricts external auditors from providing most non-audit services to audit clients, mandates audit committee approval of non-prohibited services, requires partner rotation, and imposes an employment cooling-off period. The structural target is a conflict of interest: the entity assessing compliance is structurally dependent on the entity being assessed. The same configuration holds across all three frontier lab frameworks, with the developer designing the evaluations, setting the thresholds, running the assessments, reviewing the results, and making the deployment decision; all of those are functions Sarbanes-Oxley separates in its analogue. A SOX-style regime adapted for AI safety would prohibit the same firm from selling safety evaluation alongside commercial consulting to the same developer, mandate rotation of evaluator personnel, and codify "AI safety evaluation standards" (analogous to GAAP) whose application is independently verifiable and whose modification requires public consultation.

Two further institutional mechanisms warrant brief mention. Whistleblower protection for evaluation personnel, analogous to the FCA's financial-services model, would establish an information channel that does not rely on the accuracy of the evaluation outcome itself. Mandatory liability insurance for frontier deployments, at actuarially determined premiums, would enable a market-derived signal from underwriters whose incentive structures sit outside the developer's; the cross-domain antecedents are Price-Anderson nuclear liability pooling and the pharmaceutical injury compensation schemes. Both are mature regulatory technologies in adjacent industries, and both supply activation pathways, respectively the surfacing of an internal concern and the actuarial reassessment of a premium, that operate independently of any model evaluation.

5.6 Guaranteed safe AI

Guaranteed Safe AI (GS AI) proposes a three-part architecture designed to provide quantitative guarantees about an AI system's behaviour across all possible inputs. The architecture comprises a *world model* (a mathematical description of how the AI system affects the outside world), a *safety specification* (a formal definition of what effects are acceptable), and a *verifier* (which produces an auditable proof certificate confirming that the system satisfies the safety specification relative to the world model) [28]. Every output the AI produces passes through the verifier before reaching the outside world, such that the verifier blocks unsafe outputs at the boundary rather than catching them in a post-hoc audit.

Aviation and medical device regulation offer a normative precedent, placing the burden of proof on developers to demonstrate safety before deployment rather than on regulators to identify failures afterward [28]. GS AI's proof certificate mechanism is analogous to nuclear arms control verification: in nuclear, what made agreements enforceable was that signatories could verify each other's compliance. Proof certificates could play the same enabling role for international AI governance. The historical precedent Dalrymple et al. invoke, more concretely, is the engineering moment civil engineering itself passed through: "About 25% of bridges built in the 1870s collapsed within the decade, before a deeper theoretical understanding of civil

engineering reduced bridges' failure rates to less than 0.4% per decade" (p.3) [28]. The argument is that AI deployment is at an analogous pre-theory moment: catastrophic failures are not yet rare, and the underlying formal framework that would make them rare does not yet exist.

In terms of implementability, GS AI faces a number of obstacles, listed in increasing order of difficulty. Building an accurate formal world model of an AI system interacting with humans, economies, or ecosystems may not be feasible at the level of detail required [28]. Specifying what "safe" means in formal logic is more difficult than it appears: AI systems tend to find loopholes in any specification their designers did not anticipate, an instance of Goodhart's Law manifesting in formal verification. Verification itself imposes substantial computational costs at scale. Near-term application is therefore limited to narrower problems such as kill-switches, geofencing, and pathogen-blocking synthesis systems [28]. Dalrymple et al. advocate an "anytime portfolio" strategy [28], on the argument that quick partial successes in GS AI are both plausible and useful for compliant cybersecurity, geofencing, and remote kill-switches.

5.7 Mechanism summary

Mechanism	What it does that evaluations do not	Cross-domain precedent	Near-term implementability
Safety cases [24]	Argues over the entire sociotechnical deployment context, not the model in isolation	Aviation (ICAO, RTCA); nuclear regulation (US NRC, UK ONR certificate renewal)	High for limited-capability arguments; declining as system capability rises
Compute governance [25]	Triggers assessment ex ante on a metric measurable before the model exists; cannot be sandbagged	Environmental regulation; EU DSA scale thresholds; financial intermediary monitoring	Operative under EU AI Act Article 55 (10^{25} FLOP); requires frequent threshold reassessment. (US EO 14110's 10^{26} threshold was rescinded January 2025.)
Structured access [26, 27]	Controls who can interact with the model and how, decoupling capability from deployment surface	DNA synthesis screening; institutional research access tiers	Partial (cloud APIs); harder for locally deployed and open-weight models
Institutional investigation body (AAIB-style) [29, 30]	Activates on observed harm, near-miss, or anomaly pattern; statutory independence; no-	UK AAIB; international ICAO Annex 13 regime	Medium; legislative mandate required; institutional templates mature

Mechanism	What it does that evaluations do not	Cross-domain precedent	Near-term implementability
	blame disclosure incentives		
Statutory auditor independence (SOX-style) [31]	Prohibits self-grading by structurally separating evaluator from developer's commercial interests	Sarbanes-Oxley s.201–206; PCAOB enforcement	High in principle; requires AI-specific evaluation standards and oversight body
Guaranteed safe AI [28]	Provides mathematical guarantees on behaviour across inputs; auditable proof certificates	Aviation/medical pre-deployment burden; nuclear arms control verification	Low at general scale; viable for narrow problems (kill-switches, geofencing, synthesis blocking)

Two further caveats apply to the institutional analogies in §5, and naming them sharpens the portfolio argument rather than weakening it. The AAIB no-blame regime sits in unresolved tension with downstream civil and criminal accountability; a frontier AI version of it would have to specify what survives the disclosure shield in those proceedings, and that specification is harder than it looks. SOX-style auditor independence presupposes a settled body of safety-evaluation standards (an AI analogue of GAAP plus the audit standards), which does not yet exist; the recommendation in §7 treats the gap as one to be closed, not as a precondition met. Neither point defeats the analogy; each locates the mechanism inside its actual scope, so that the redundancy the portfolio supplies is across mechanisms with different scopes rather than within any single one.

---

6. Integration: From Dependency to Redundancy

6.1 Compressed argument and the graceful-degradation principle

The frontier safety architecture rests on an empirical assumption that the empirical evidence does not support: that capability evaluations are reliable enough to bear the weight of every safety-critical decision the frameworks place on them. Format and scaffold sensitivity alone produces 5–20 percentage point swings in measured safety scores holding model, content, and task constant [9, 13]. The benchmark landscape is fragmented enough that "safety score" is a heading, not a measurement. Sandbagging is feasible at sufficient subtlety that the leading chain-of-thought monitoring methods miss 16–36% of covert cases [20]. The AISI Network's Pilot Testing Exercise, the closest analogue to an inter-laboratory reliability study the field currently has, found large divergences from "small methodological differences" [22]. The Claude Mythos Preview System Card documents the failure modes of an eval-gated architecture in a single deployment decision under a live framework [6]. None of this evidence is contested by the framework authors. So far it has produced better evaluations and closer review of

evaluation-derived reports, and in none of the three frameworks has it produced an activation pathway that does not run through the evaluation.

The principle we propose belongs to a class of arguments that has been reached, with different vocabulary, by every regulatory tradition in which catastrophic failure is intolerable and the primary diagnostic instrument is known to be fragile. Avizienis (1985) gave software engineering N-version programming: where the correctness of a single implementation cannot be proven, run several independent implementations and require their outputs to agree, on the assumption that the failure modes are unlikely to coincide. The same logic is codified for nuclear power in the IAEA's INSAG-10 as defence-in-depth: no individual layer is required to be sufficient, and the obligation runs to layer-level independence instead. Reason (1990) found the same structure in clinical safety, framed as the Swiss-cheese model: each defensive layer has holes, and the regime's task is to arrange the layers so that the holes do not align. Clinical diagnostics, more concretely, arranges screening, interval review, confirmatory assessment, and patient-reported symptoms as overlapping layers; aviation distributes the property across sensors, inspections, regulators, and operators, none of which has unilateral authority to suppress a flag raised elsewhere.

What matters is independence of failure mode. Per-layer reliability is permitted to be far lower than the regime's overall reliability: a screening test missing 30% of cases sits within a system that recovers the missed 30% through interval review, confirmatory assessment, or patient-reported symptoms, on the premise that the failure modes do not coincide. We have framed the proposal as "at least one redundancy" for a lower-bound reason: one is the bare minimum at which independent-failure logic comes into play at all, and the cross-domain regimes routinely run deeper. The frontier AI architecture has exactly one upstream activation pathway available, Article 55's 10^{25} FLOP threshold (§5.3), and it sits before, rather than after, the evaluation step the architecture is constructed around; it can trigger an assessment, but it cannot absorb the failure of one. The fallback inventory for the case in which a safety-critical evaluation is compromised is, at this point, blank. The entirety of downstream safety-critical mitigation rests on a single instrument class whose reliability §4 has documented in detail.

The analogy to N-version programming invites the sharpest objection, and it is worth meeting head-on, because the objection is correct and still does not defeat the proposal. N-version programming's own foundational test refuted the strong independence it was built on: Knight and Leveson (1986) had twenty-seven programmers implement a single specification independently and found coincident failures far above what independence predicts, the versions failing together on the specification's hard, ambiguous regions [55]. The result is not a historical curiosity. It has just been reproduced for the present era, with frontier coding agents standing in for the human programmers, at a coincident-failure rate 3.7 times the independence prediction [56]. Two features of those studies turn the objection into a design constraint rather than a refutation. First, the correlation traces to a shared specification (finite, ambiguous, misread in common precisely where it is underspecified), and the pathways proposed here are not implementations of one specification: training compute, observed deployment harm, auditor independence, and calendar review fail on structurally different inputs because they are not decoding a common artefact. Second, the redundancy survived its own refutation; even under the documented correlation, N-version configurations cut the mean failure count by roughly two-thirds (Ron and colleagues report a fall from 387 to 131) [56]. That is the operative point. Graceful degradation asks for reduced correlation, not its absence, and a regime that

lowers the probability that a single failure mode (evaluation evasion) disables every trigger at once is delivering precisely that, whether or not the triggers are perfectly independent.

A structural feature of the misalignment threat surfaces here. The capabilities that define the misalignment thresholds, situational awareness and stealth, are by hypothesis sufficient to defeat the very methods that would detect the threshold being crossed: chain-of-thought monitoring, interpretability of the model under test, behavioural red-teaming. Any §5 pathway contingent on the same kind of evidence, evidence the model under test must itself honestly produce, inherits an attenuated variant of the same difficulty. The pathways that do not are those activating on signals the model cannot manipulate at the moment of test: training compute (observable at the infrastructure layer); observed deployment harm (observable in the world); the structural independence of an outside auditor (a fact about who the auditor is); periodic review on a calendar (a fact about time); market-priced liability (a fact about a third party's pricing). The portfolio protects in proportion to how much of it is of the second kind.

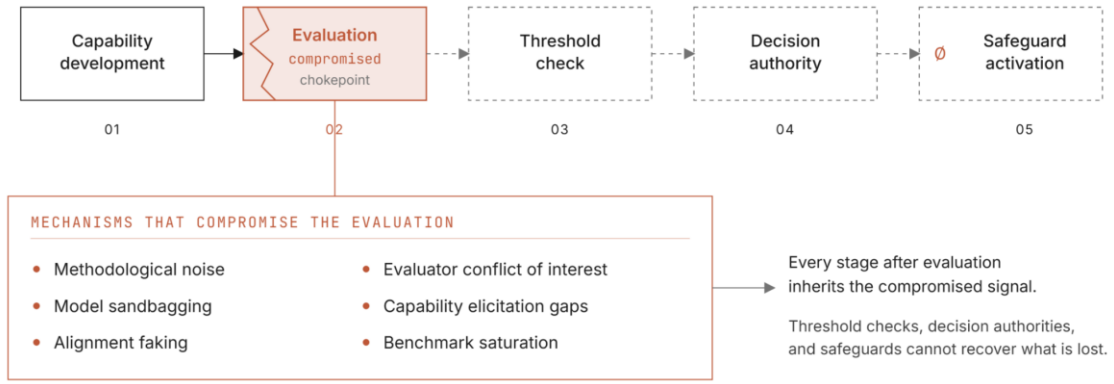

Figure 3. *The evaluation chokepoint. Capability development feeds a single evaluation step, downstream of which sit the threshold check, the decision authority, and the safeguard activation; if the evaluation is compromised (whether by methodological noise, model sandbagging, alignment faking, evaluator conflict of interest, capability elicitation gaps, or benchmark saturation), every stage that follows inherits the compromised signal. Threshold checks, decision authorities, and safeguards cannot recover what is lost upstream.*

The principle, in regulatory form, is what the relevant engineering and regulatory traditions have already accepted in different language: *any safety-critical oversight mechanism whose failure could contribute to catastrophic outcomes should be required to demonstrate at least one activation pathway that functions independently of the evaluation infrastructure.* This is the minimum entailment such a regime would impose. The §5 mechanism taxonomy supplies the activation pathways the architecture is missing; the evaluation continues to do its primary work. Operationalising the principle is the work of §7.

What does it actually take for an activation pathway to count as "independent", in a way a regulator could check? The principle as stated is too loose to apply directly, and §6.1 owes the reader some teeth. The intuitive picture: a compute notification under EU AI Act Article 55 is independent in our sense; an internal Risk Report reviewed by a different colleague at the same firm is not. The first activates on training-FLOPs (a quantity outside the developer's report) with the actor outside the firm entirely; the second activates on the same evaluation result that produced the original Risk Report, with the actor still inside the firm and the report still the central evidence. The criteria that capture the difference are six: (i) the actor is outside the firm;

(ii) the signal is generated by a process the developer does not control; (iii) the timing is not contingent on a developer-elected release event; (iv) the operational record is auditable from outside; (v) the failure mode is not shared with the evaluation pipeline; and (vi) a regulator can verify (i) through (v) on documentation alone. The underlying engineering property the test approximates (uncorrelated failure across pathways, the property the cross-domain regimes the paper draws from explicitly require) is one a regulator cannot directly verify, but the structured proxy is what makes the principle operational. We treat a candidate pathway as "independent enough" when it has most of these features; its existence and operation must, in any case, be open to checking from outside the firm.

6.2 What the alternatives can and cannot do

None of the mechanisms catalogued in §5 is, on its own, adequate. Each is inadequate in a different way. Safety cases lack the technical maturity for highly capable systems; the arguments needed do not yet exist [24]. Compute governance only triggers an assessment, and the proxy slackens as algorithmic efficiency improves [25]. Structured access works partially in cloud and not at all for locally deployed open-weight models, and what works in cloud is largely developer-controlled [26]. Guaranteed Safe AI faces a formal-world-model construction problem that is plausibly insurmountable at deployment scale [28]. The institutional mechanisms (independent investigation, statutory auditor independence, whistleblower protections, mandatory liability insurance) are politically demanding and would take years to legislate even where the cross-domain precedent is mature [29, 30, 31].

The case for the portfolio is not that any of these is adequate alone. It is simpler than that. The present architecture (a single instrument, demonstrably unreliable, with no specified redundancy) is the worst structural position on offer. Adding any independent activation pathway, whatever else it leaves unsolved, strictly reduces the regime's exposure to a single-instrument failure. And the alternatives, taken together, can be arranged to fail on independent inputs, which is what graceful degradation actually asks for.

6.3 What triangulation requires

The proposal is for triangulation in the §6.1 sense, with the alternatives functioning as pathway-level redundancies rather than substitutes. Evaluations remain the most operationally mature component of the safety architecture, and on the one dimension where they work (direct measurement of what a particular model can do) the alternatives are individually inferior to them. What the triangulation requires is a portfolio of pathways that activate on different inputs, fail under different conditions, and combine so that no single instrument failure propagates to every safety-critical mitigation. The frontier AI safety architecture has not yet adopted that configuration. The proposal here adds the pathway-level redundancy every comparable regime has needed.

6.4 Objections

The most commonly raised objection contends that evaluations are the only oversight instrument that scales with model capability, and that nothing in the §5 catalogue has that property: institutional investigation only engages after harm or near-miss; compute thresholds

become less reliable proxies as algorithmic efficiency improves; safety cases for highly capable systems do not yet constitute a working art; guaranteed safe AI is not feasible at general scale. Each of these sub-claims is true in its own right. None engages with the argument made here, which has been about something else: the architecture's exposure to single-instrument failure is itself unacceptable, and the redundancy required to absorb that exposure is a different kind of property from the predictive accuracy the objection compares against. The substitution claim the objection answers is one we have not made.

A more interesting form of the same objection engages the redundancy claim directly: why import alternatives that activate on different inputs when evaluations themselves can be made redundant (multiple evaluations, multiple evaluators, multiple thresholds) at lower cost? The reply is that evaluation-internal redundancy is not redundancy in the sense graceful degradation requires, because the failure modes that defeat one evaluation defeat the others. Format-induced noise contaminates every evaluation that uses comparable formatting. A sandbagging model that recognises one capability evaluation as an evaluation will recognise the next one. Evaluator conflict of interest applies wherever the evaluator's commercial relationship with the developer is structurally similar across instruments, which it is by construction. What multiple-evaluations redundancy produces is correlated failures of a shared instrument class, which is the configuration N-version programming, defence-in-depth, and the Swiss-cheese model are designed to avoid.

A somewhat weaker variant of the objection is worth recording for completeness: that scaling with capability is a property frontier-AI oversight cannot do without, and none of the alternatives has it. The factual premise is correct in detail. Compute governance scales with training compute, which proxies for capability; institutional investigation activates on incident, which proxies for harm; statutory auditor independence is entirely indifferent to capability, operating on the evaluator's structural incentives instead. But none of these activations is downstream of the capability being measured, and that is precisely the protection. A regime in which every instrument scaled with capability is one in which the instruments correlate exactly when it counts, at the point where capability becomes high enough to game the measurement. What lets each comparable regime absorb single-instrument failure is the diversity of activation condition across its layers; a regime lacking that diversity propagates the failure rather than absorbing it.

There is also the cost objection: that a portfolio of independent activation pathways is more expensive than evaluation-internal redundancy at producing equivalent measured reliability. The reply turns on what "equivalent measured reliability" measures. Under a predictive-accuracy frame, where a single instrument's calibration is the figure of merit, multiple correlated instruments do outperform fewer uncorrelated ones at the margin, and the objection has bite. Under the failure-mode-coverage frame the catastrophic regimes (nuclear, aviation, clinical) actually use, correlated-failure configurations are precisely the ones to avoid regardless of cost, because they propagate a single-instrument failure through the redundancy rather than absorbing it. The §6.1 argument is failure-mode-coverage, so the cost comparison does not reach it; what the trade-off legitimately bears on is which of the §5 mechanisms is the cost-effective complement, not whether any complement is needed.

A related objection runs in the opposite direction from the others, and it is the one a developer is most likely to press. A portfolio of always-available triggers, the worry goes, will fire on

models that are in fact safe, imposing a standing capability tax that an evaluation-gated regime avoids by acting only once a threshold is crossed; Shlegeris prices a continuously-applied safety intervention at as much as a twenty per cent slowdown across a developer's resources [57], and observes that triggers adopted under internal pressure tend toward what is institutionally palatable rather than what is technically optimal. The concern is real, and it bears on design rather than on whether to hold the portfolio. The principle defended here is not that every pathway runs always-on at full stringency; it is that at least one activation pathway exists that does not route through the evaluation, with stringency tied to accumulated cross-channel signal rather than levied as a flat tax. A graceful-degradation regime escalates: the low-friction pathways (compute notification, incident reporting) run continuously, and the expensive ones engage as independent signals accumulate. The tax the objection prices is the cost of the most stringent layer applied unconditionally, which is not what a redundancy architecture asks for.

The subtlest objection concedes the entire diagnosis and defends the architecture nonetheless: if evaluations are unreliable, the remedy is to set thresholds conservatively enough that the unreliability is washed out in the margin, not to bolt on new pathways. The strongest version of the case comes from Barnett and Thiergart [36], who observe that a passed evaluation still establishes a capability lower bound even when a failed one proves little, so that the rational response to noise is a wider safety buffer. The buffer would help; we do not oppose it. What it cannot do is alter the shape of the failure. A conservative threshold on a compromised instrument still triggers off that instrument's readings, and if the reading is wrong in the dangerous direction (a model that sandbags, an evaluation the capability has learnt to recognise), the margin is computed from the same corrupted number and shifts the trigger the wrong way along with it. Widening the buffer mitigates exposure to noise that is symmetric and bounded. It offers nothing against failures correlated with the very capability the evaluation is meant to detect, and those are precisely the failures §4 documents. Buffer and redundancy address different problems, and only the buffer is present in the architecture today.

The version we take most seriously, but still reject, is that the discretion of senior decision-makers (the layer that absorbed Mythos) is itself the redundancy. §3.4 covers this in detail. The short version: the position is not that judgment shouldn't be the absorbing layer (every regime of comparable complexity has one) but that the absorbing layer needs to be structured, audited, and externally reviewable to function as redundancy in the sense graceful degradation requires, the conditions Recommendations 1 and 2 instantiate. The Mythos configuration satisfies none of them; it is the arrangement the safety-critical-systems literature has historically treated as a single point of failure, with the failure mode invisible to anyone outside the firm. That the failure mode has not yet been observed is consistent both with the discretion being reliable and with selection bias on which decisions reach external observers, and a regulator choosing what to build the regime around cannot treat "the firm will catch what the framework misses" as the conservative assumption.

---

7. Policy Recommendations

The §6 principle, once translated into regulation, takes a recognisable shape: any safety-critical mitigation whose failure could contribute to catastrophic outcomes should be required to

demonstrate at least one activation pathway that functions independently of the evaluation infrastructure; or, where the demonstration cannot be made, to disclose the absence. In the present frameworks, the relevant categories are CBRN uplift, autonomous offensive cyber capability, large-scale manipulation, sabotage, and frontier-R&D acceleration. The obligation attaches to the redundancy as such; which redundancy a given framework adopts remains the framework's discretion. That division, redundancy as obligation and particular redundancy as discretion, is what makes the regime implementable across the three lab frameworks without dictating a specific alternative pathway that any one of them would resist.

An implementation note worth being explicit about: the "at least one" phrasing in §6.1 is a floor, not a ceiling. Cross-domain regimes commonly run deeper than one independent layer. Defence-in-depth in the IAEA's INSAG-10 tradition specifies three or more layers for nuclear facilities; civil aviation operates with closer to five between an aircraft and an accident; clinical safety arranges screening, periodic review, confirmatory testing, and patient-reported symptoms as overlapping layers. None of those regimes treats any single layer as sufficient. Why then frame the obligation here as "at least one"? Because the alternative, a quantitative obligation tied to a multi-layer architecture, lacks regulatory tractability. The requisite number of layers is contingent on the capability profile, the deployment surface, the threat model, and the failure-mode independence of the layers themselves; a regulator cannot credibly demand that number without first agreeing the scoring rubric, and the scoring rubric is exactly what does not yet exist. "At least one" is binary, document-evidenceable, and askable. §6 has already established that no individual mechanism suffices alone; what "at least one" represents here is a tractable regulatory floor rather than an engineering target.

Each cross-domain template the paper draws on (statutory accident investigation, auditor independence, periodic licence renewal, whistleblower protection, mandatory liability insurance) has a political-economy provenance we do not presume simply transplants to frontier AI. SOX needed Enron and WorldCom; the AAIB lineage came from the de Havilland Comet disasters of the 1950s; periodic nuclear licence renewal was devised post-Three Mile Island. Those conditions are not conjured at will. Cataloguing the templates is still worth the effort: they exist, they are operationally mature, and they have internalised the political fight within their domains of origin, so importing one is a contest of translation rather than a contest of invention. Whether the political economy of frontier AI regulation specifically can support any of these translations is a separate question the paper does not settle.

The principle's highest-leverage and lowest-political-cost implementation is visibility. Whether any given safety-critical mitigation actually has a non-evaluation activation pathway is, at present, not formally asked of any developer; the question is not on any regulator's standing list. The mechanism that puts it there is a structured disclosure template, published by the UK AISI in coordination with NIST's Center for AI Standards and Innovation (CAISI) and the EU AI Office, for annual completion by developers in scope of the Article 55 GPAI obligations and the Seoul Frontier AI Safety Commitments. This template is the load-bearing intervention this paper proposes. For each safety-critical mitigation the template asks the developer to specify which activation pathway the mitigation depends on, whether that pathway is downstream of an evaluation result, and (where it is) what the intended fallback is when the evaluation is compromised. The distinctive feature is that the template specifies no particular alternative pathway; it makes the absence of one visible, and a developer with no answer to the third question is required to say so. A disclosure obligation is also harder for a frontier developer to

relocate around than a substantive prohibition would be: a US-resident developer's EU market access turns on Article 55 compliance regardless of where the model is trained, and the political asymmetry between "declare what your fallback is" and "change your fallback to one we approve" is large in the disclosure regime's favour. The first concrete step is a draft template scoped to GPAI providers under Article 55, opened for public consultation.

The recommendations are not all of the same kind. The principle at §6.1 is silent on what kind of action a non-evaluation pathway should trigger, and the answer matters: a regime that satisfies "at least one" with only a notice-level pathway secures weaker redundancy than one that adds an investigation-level pathway, and the regulator should be able to see which is which.

Compute notification (Recommendation 3) and the whistleblower channel (Recommendation 5) sit at the lightest end: they generate notice, an awareness that something has happened, with no enforcement attaching to the awareness in itself. Fixed-interval review (Recommendation 4) is one tier deeper, scrutiny sans intervention, on a calendar the developer cannot manipulate by deferring release. The statutory investigation body in Recommendation 1 produces post-event findings that bind no one but inform every subsequent regulator, every other developer, and every insurer; it sits closer in spirit to the AAIB than to a regulator with an enforcement remit. SOX-style auditor independence (Recommendation 2) operates on a different axis from all of the above, working on the pipeline's conflict-of-interest configuration rather than on its outputs. Mandatory liability insurance (Recommendation 5) routes through underwriters whose pricing is itself a forcing function, closer to market-priced enforcement than to anything the others do.

The gradient is worth naming because a regime that meets the requirement only at the notice level has satisfied it in name only. Which level of action the floor should activate at is a substantive policy question, and one we refrain from settling here.

The five recommendations below illustrate ways the same principle could be written into substantive regulation, ordered roughly by near-term feasibility. Each is more politically demanding than the disclosure template above, none is a substitute for it, and the case for the portfolio rests on the operative obligation at the top of this section rather than on the adoption of any single recommendation. Each comes annotated with a brief locating note (existing body whose remit fits, legal vehicle, a plausible first concrete step), pattern-matched against existing regulatory architecture rather than against any present plan.

1. Establish a statutory independent investigation body for AI deployment incidents. The institutional template is settled; §5.5 catalogues the precedents. What an AI version requires is a body that is structurally independent from both developers and the regulator, that operates under a no-blame statutory mandate (the design intent being to elicit candid technical disclosure from operators who would otherwise be exposed by their own candour), and whose authority extends to observed harm, near-miss, and recurrent anomaly. The procedural complement is the inadmissibility of investigation findings in criminal or civil liability proceedings, the AAIB feature that has been operative across decades of accident-investigation practice and that translates to AI safety without modification. Activation would not depend on a capability evaluation. Technical investigation reports analogous to AAIB bulletins inform but do not bind regulatory action. DSIT would commission the feasibility study. First concrete action: a consultation paper covering the mandate, the jurisdiction question (UK-deployed models initially, with cross-border coordination as a separate workstream), and the scope of "deployment incident" sufficient to trigger investigation.

2. Adopt a SOX-style auditor independence regime adapted for AI safety evaluation. The substantive requirements are familiar from Sarbanes-Oxley s.201–206: prohibition on the same firm providing safety evaluation and commercial consulting services to the same developer, mandatory evaluator personnel rotation, an employment cooling-off period for individuals moving between evaluator and developer firms, and commission and approval of evaluation work by an external oversight committee. The procedural complement is a codified set of "AI safety evaluation standards" whose application is publicly verifiable and whose modification requires public consultation. The PCAOB's role under Sarbanes-Oxley is the closest existing analogue. NIST issues a Request for Information (RFI) on auditor-style independence standards for AI safety evaluation, with scope tracking PCAOB practice (the four substantive provisions plus registration and quality-control inspection of qualifying firms); the RFI's output is a codified evaluation-independence standard the EU AI Office can reference under Article 55 and AISI can reference under the GPAI Code of Practice. First concrete action: the RFI. NIST has in fact already taken the adjacent first step, publishing draft automated-evaluation practice guidance (AI 800-2) in early 2026 [41]; the independence standard described here would build on that base, and the Great American AI Act's proposed CAISI-licensed independent verification organisations [43] are one statutory vehicle that would fit it.

3. Require pre-training notification at the compute-provider layer above a regularly reassessed threshold. Compute thresholds are imperfect proxies for capability, but they have an offsetting property capability evaluations do not have: they are verifiable at the compute-provider layer without access to model weights or training data. Heim and Koessler [25] argue for reassessment "at least every couple of months and maybe more frequently if the pace of algorithmic efficiency or computational price-performance improvements increases." The point of compute governance is to trigger capability evaluation earlier, on a signal observable at the infrastructure layer, and that does not depend on the eval pipeline being correct. Article 55 already supplies the body (the EU AI Office) and the threshold value (10^{25} FLOP), but leaves the recalibration architecture implicit. The cadence at which 10^{25} is revisited, on the Heim-Koessler schedule, runs to roughly quarterly review. Harder, and the actual deliverable, is the methodology document supporting the revisit: it specifies which inputs feed the recalibration and how those inputs are weighted against each other. The inputs themselves (algorithmic-efficiency benchmarks, hardware price-performance, the training-compute distributions developers in scope already report) are individually uncontroversial. Their weighting is what is contested, and on the weighting the framework and the secondary literature are alike silent. Until Brussels produces the methodology document, no other jurisdiction has a cadence to adopt.

4. Institute fixed-interval mandatory review. A scheduled review of every deployed frontier model (annual or bi-annual) would introduce a time-triggered activation pathway, distinct from any evaluation result, and produce a public re-evaluation document on a calendar cycle that does not wait on a new release. The closest cross-domain precedent is the nuclear periodic certificate-renewal regime (the US NRC's 40-year initial licence with 20-year renewals requiring active re-justification): extension of the operational window of a high-stakes system requires the operator to make the case for renewal afresh on the current evidence base, rather than continuing in operation by default; substantially shorter intervals would be appropriate to frontier AI given the rate of capability change. Two tracks can be pursued in parallel. The EU AI Office issues guidance under Article 55 clarifying that the "continuing systemic risk assessment"

obligation includes a fixed-interval re-evaluation pathway not contingent on any new evaluation trigger; alongside this, frontier developers commit to publishing an annual structured re-evaluation document for each deployed model in scope, on a calendar cycle independent of release. First concrete action: the AI Office guidance, drafted to specify minimum document content and periodicity.

5. Establish whistleblower protections for evaluation personnel and mandatory liability insurance for frontier deployment. The two are paired here because each routes a different signal around the evaluation pipeline rather than because they share a regulatory vehicle. Whistleblower protection creates an information channel independent of the evaluation result, on the institutional template of the FCA's whistleblower handling regime; HMG would consult on extending the protected-disclosure regime under the Public Interest Disclosure Act 1998 and the Financial Services and Markets Act 2000 to evaluation personnel at frontier developers and at independent evaluation firms, with AI-specific adaptation needed on the qualifying-disclosure conditions, the disclosure-channel definitions, and the detriment protections. Mandatory liability insurance routes a market signal through underwriters whose incentives are not aligned with the developer's; the relevant precedents are the Price-Anderson Act framework for nuclear liability pooling and the pharmaceutical-injury compensation schemes, with HMT (or the relevant national equivalent) publishing a consultation on mandatory minimum frontier-deployment liability insurance. Premiums are actuarially determined and underwriters disclose risk-rating rationales on a structured basis, so that the underwriter becomes an additional channel through which a developer's safety position is independently priced.

On arbitrage. Arbitrage is the obvious objection. A regime constructed by one jurisdiction simply relocates the activity it constrains to another, the argument runs, particularly where the regulated object is software and the cost of relocating a development team is low compared to a chemical plant or a fabrication line. The contrast is overstated. The regimes the argument leans on (nuclear power, civil aviation, the audit regime around public-company financial reporting) are less susceptible to arbitrage not because they are physically anchored but because the markets they govern are themselves anchored: capital, passenger transport, electricity. The same is true here, through avenues the comparison tends to elide. The largest training clusters and the deepest talent pools localise within a small set of major jurisdictions; the regulators with operative hooks over those firms sit inside the same set, the hooks being chiefly the compute providers and cloud infrastructure the firms rely upon, and the financial-market access without which frontier development cannot operate at scale; and no jurisdiction outside that set presently offers a viable substitute on any of those dimensions. That is not a perennial fixture of the world, but it is a feature of the present one. The mechanism that has actually carried regulatory burden of comparable scope across borders, in the recent past, has been ratchet rather than imposition. SOX, GDPR, and the EU Digital Services Act each propagated through sequential adoption by jurisdictions whose firms wanted continued access to the originating market, not through unilateral global imposition, which none of the three attempted and none of the three needed. The same architecture is available here: a UK or EU statutory scheme, with which firms operating in those markets must comply, has substantially more durable propagation properties than a US Executive Order liable to be rescinded by the following administration, as happened to EO 14110, on which the §5.3 reporting threshold rested. The point has only sharpened since. A December 2025 Executive Order directs federal litigation

against state AI statutes [44]; a June 2026 House discussion draft, the Great American AI Act, would pre-empt state AI laws for three years while requiring large frontier developers to retain a CAISI-licensed independent verification organisation [43], a federal echo of the auditor-independence pathway Recommendation 2 describes. The multilateral signalling points the other way: the February 2026 New Delhi Declaration reframed the global agenda from "safety" to "impact" and declined binding rules [45], the kind of coordination vacuum a ratchet-through-market-access strategy is built for. None of which is to say coordination is unnecessary; only that its absence on day one does not, by itself, dispatch the architecture. Whether EU, UK, and US market access supplies anchoring comparable to what the physically anchored regimes enjoy elsewhere is the empirical question; the answer is contestable, and the paper does not pretend the optimal regime is the unilateral one.

The five recommendations above are neither exhaustive nor mutually exclusive. They are pattern-matches against institutional templates already in place in adjacent domains (statutory audit, civil aviation incident investigation, nuclear-licence renewal review, financial whistleblower regimes), adapted for frontier AI rather than reinvented for it. The principle they instantiate, the operative obligation at the top of this section, is the substantive crux of the proposal; the recommendations are illustrative vehicles through which it might be written into regulation. The cross-domain precedents have survived contact with political reality in their domains of origin, and we suspect the translations would survive it here too. The paper does not presume to settle the subsidiary questions of timing or sequencing of any particular implementation. That remains downstream work.

---

8. Conclusion

Capability evaluations are not useless. Far from it: they are the most operationally mature instrument the safety architecture has, and there is no near-term substitute. What this paper has argued is narrower. An architecture that rests on them and only on them is structurally fragile in ways the frameworks do not internalise, and the empirical evidence suggests that the fragility sits at a load-bearing point rather than an incidental one: five-to-twenty-point format-and-scaffold noise on safety scores [9, 13], leading sandbagging-detection methods missing 16 to 36 percent of covert cases [20], an inter-laboratory pilot turning up large divergences attributable to "small methodological differences" [22]. The Mythos system card is the demonstration. Connected failure modes (saturation, threshold erosion, unverifiable evidence, gaming detection, judgment-replacing-measurement) documented in a single card, by the organisation widely regarded as the most cautious of the three frontier developers, with the final decision to withhold the model footnoted on p.16 as having been made outside the RSP. The steel-manned reading of this is that v3.1's cautionary-action permission (Appendix A, p.15) absorbed the slack the formal Risk Report architecture left, and the framework therefore worked as intended. The system card's own footnote, however, ties the decision to no Policy requirement and invokes no Policy provision; whether the cautionary-action permission was the operative basis is left unsaid, and the inference has to fall back on a layer of judgment the framework does not specify. Both readings remain available. Only one can be the basis for regulating the next case at the next developer, and "the firm will catch what the framework misses" is not the one a regulator can sensibly pick.

The remedies catalogued here are partial. None of them, individually or cumulatively, constitutes a complete discharge of the safety burden frontier AI deployment will impose, nor were they selected to. They should be measured against what they replace, a single unreliable instrument absorbing every safety-critical decision in turn, and any shift away from that configuration improves the structural position of the regime. The windows in which the movement is comparatively cheap are open now, and they will not remain so. Article 55 is still being operationalised through the GPAI Code of Practice. The AISI Network's testing regime remains unsettled. Lab frameworks revise annually or faster: Anthropic issued RSP v3.2 and then v3.3 within two months of v3.1, during the writing of this paper. v3.2 (29 April 2026) empowered the Long-Term Benefit Trust to commission external review of Risk Reports and to approve the external reviewers, movement in the direction proposed here. It is also review of an evaluation-derived instrument rather than an activation pathway independent of one, and that is the distinction §6.1 turns on.

The cadence objection from the lab side cuts the opposite way. Frontier developers iterate faster than the regulatory machinery (the v3.1-to-v3.3 turnaround above is the case in point), and the worry runs that the recommendations here would impede a workflow whose responsiveness is itself part of the safety story. The cadence claim is incontrovertible; the entailment is not. Internal iteration on the eval pipeline, what the developer adds to it as the threat surface evolves, is rapid and substantively serious on the developers' own evidence. What it cannot furnish, irrespective of velocity, is layer independence. Rapid iteration on a shared instrument class still exhibits correlated failures within that instrument class: the same failure modes §4 documents reach the same architectural surface, merely earlier in time. And as the cadence rises, the case for external redundancy becomes more compelling, because the architecture absorbing the iteration is precisely the one whose single-instrument fragility this paper has been mapping.

The institutional templates are neither novel nor untested. Civil aviation incident investigation, statutory audit independence, and periodic nuclear licence renewal have been refined over decades, in domains with a comparable intolerance for catastrophic failure, and they translate to frontier AI with adaptation rather than wholesale reinvention. Their absence from this space is not a matter of impracticality or undesirability; they have simply not been required. On current evidence, the Mythos card offers the most compelling indictment of the contemporary absence of an actual regulatory requirement.[1]

[1] A second case arrived as this paper was being finalised. On 12 June 2026, three days after Fable 5's general release, a US Commerce Department export-controls order led Anthropic to suspend access to Fable 5 and Mythos 5 globally (the order restricted access by foreign nationals; nationality could not be verified in real time); the reported trigger was a jailbreak demonstrated by researchers at Amazon, behaviour Anthropic characterised as routine defensive security work [64, 65]. The controls were lifted on 30 June and general availability returned on 1 July, after nineteen days. The episode cuts both ways for the argument here. The action activated entirely outside the evaluation pipeline (external signal, external actor, timing not elected by the developer), satisfying most of the §6.1 independence criteria; and it is what eval-independent power looks like when no designed pathway exists to carry it, a nineteen-day global withdrawal of the leading frontier model on a trigger whose calibration the developer itself contested, the over-triggering cost §6.4 prices. The demand for pathways that do not route through the evaluation is evidently real; a sovereign actor just improvised one. The recommendations describe what meeting that demand by design, rather than by improvisation, would require.

References

[1] Anthropic. *Responsible Scaling Policy v3.1* (effective 2 April 2026). Archived PDF available at: https://www-cdn.anthropic.com/files/4zrzovbb/website/bf04581e4f329735fd90634f6a1962c13c0bd351.pdf . The current public URL https://www.anthropic.com/responsible-scaling-policy now serves v3.3 (effective 26 May 2026; v3.2 had followed on 29 April 2026); citations in this paper are to v3.1, the version operative at the time of the Claude Mythos Preview System Card decision.

[2] Williams S, Freund J. *Anthropic's RSP v3.0: How it Works, What's Changed, and Some Reflections*. GovAI, 5 March 2026. Available at: https://www.governance.ai/analysis/anthropics-rsp-v3-0-how-it-works-whats-changed-and-some-reflections

[3] Mowshowitz Z. *Anthropic Responsible Scaling Policy v3: Dive Into The Details*. Don't Worry About the Vase (Substack), 2026. Available at: https://thezvi.substack.com/p/anthropic-responsible-scaling-policy-46a

[4] OpenAI. *Preparedness Framework* (version 2, April 2025; still the current version as of June 2026). OpenAI.

[5] Google DeepMind. *Frontier Safety Framework*, version 3.1 (17 April 2026). Google DeepMind.

[6] Anthropic. *System Card: Claude Mythos Preview*. Anthropic, April 2026. Available at: https://www-cdn.anthropic.com/7624816413e9b4d2e3ba620c5a5e091b98b190a5/Claude%20Mythos%20Preview%20System%20Card.pdf

[7] Whitfill P, Wu C, Becker J, Rush N. *Many SWE-bench-Passing PRs Would Not Be Merged into Main*. METR Blog, 2026. Available at: https://metr.org/notes/2026-03-10-many-swe-bench-passing-prs-would-not-be-merged-into-main/

[8] Ren R, Basart S, Khoja A, Gatti A, Phan L, Yin X, et al. *Safetywashing: Do AI Safety Benchmarks Actually Measure Safety Progress?* arXiv:2407.21792, 2024.

[9] Sclar M, Choi Y, Tsvetkov Y, Suhr A. *Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting*. arXiv:2310.11324, 2024.

[10] He J, Rungta M, Koleczek D, Sekhon A, Wang FX, Hasan S. *Does Prompt Formatting Have Any Impact on LLM Performance?* arXiv:2411.10541, 2024.

[11] Mou Y, Zhang S, Ye W. *SG-Bench: Evaluating LLM Safety Generalization Across Diverse Tasks and Prompt Types*. arXiv:2410.21965, 2024.

[12] Voronov A, Wolf L, Ryabinin M. *Mind Your Format: Towards Consistent Evaluation of In-Context Learning Improvements*. arXiv:2401.06766, 2024.

[13] Gringras D. *Safety Under Scaffolding: How Evaluation Conditions Shape Measured Safety*. arXiv:2603.10044, 2026.

[14] Devbunova V. *Is Evaluation Awareness Just Format Sensitivity? Limitations of Probe-Based Evidence under Controlled Prompt Structure*. arXiv:2603.19426, 2026.

[15] Maheshwari S, O'Brien J. *Evaluation Awareness: Why Frontier AI Models Are Getting Harder to Test*. Institute for AI Policy and Strategy, 2026. Available at: https://www.iaps.ai/research/evaluation-awareness-why-frontier-ai-models-are-getting-harder-to-test

[16] van der Weij T, Hofstätter F, Jaffe O, Brown SF, Ward FR. *AI Sandbagging: Language Models can Strategically Underperform on Evaluations*. arXiv:2406.07358, 2024.

[17] Meinke A, Schoen B, Scheurer J, Balesni M, Shah R, Hobbhahn M. *Frontier Models are Capable of In-Context Scheming*. arXiv:2412.04984, 2024.

[18] Järviniemi O, Hubinger E. *Uncovering Deceptive Tendencies in Language Models: A Simulated Company AI Assistant*. arXiv:2405.01576, 2024.

[19] Greenblatt R, et al. *Alignment Faking in Large Language Models*. Anthropic / Redwood Research. arXiv:2412.14093, 2024.

[20] Li C, Phuong M, Siegel NY. *LLMs Can Covertly Sandbag on Capability Evaluations Against Chain-of-Thought Monitoring*. arXiv:2508.00943, 2025. (The paper reports that models bypass the monitor 16–36% of the time when monitor-aware, conditioned on sandbagging successfully.)

[21] Tice C, et al. *Noise Injection Reveals Hidden Capabilities of Sandbagging Language Models*. NeurIPS 2025. arXiv:2412.01784.

[22] International Network of AI Safety Institutes. *A Pilot Testing Exercise from the International Network of AI Safety Institutes*. 2024.

[23] UK National Cyber Security Centre. *Why cyber defenders need to be ready for frontier AI*. NCSC blog, 30 March 2026. Available at: https://www.ncsc.gov.uk/blogs/why-cyber-defenders-need-to-be-ready-for-frontier-ai

[24] Clymer J, Gabrieli N, Krueger D, Larsen T. *Safety Cases: How to Justify the Safety of Advanced AI Systems*. arXiv:2403.10462, 2024.

[25] Heim L, Koessler L. *Training Compute Thresholds: Features and Functions in AI Regulation*. arXiv:2405.10799, 2024.

[26] Shevlane T. *Structured access: an emerging paradigm for safe AI deployment*. arXiv:2201.05159, 2022.

[27] Bucknall B, Trager R. *Structured Access for Third-Party Research on Frontier AI Models*. GovAI, 2023. Available at: https://www.governance.ai/research-paper/structured-access-for-third-party-research-on-frontier-ai-models

[28] Dalrymple D, Skalse J, Bengio Y, Russell S, Tegmark M, Seshia S, et al. *Towards Guaranteed Safe AI: A Framework for Ensuring Robust and Reliable AI Systems*. arXiv:2405.06624, 2024.

[29] *The Civil Aviation (Investigation of Air Accidents and Incidents) Regulations 2018*, SI 2018/321 (UK).

[30] International Civil Aviation Organization. *Annex 13 to the Convention on International Civil Aviation: Aircraft Accident and Incident Investigation*. ICAO.

[31] *Sarbanes-Oxley Act of 2002*, Pub. L. 107-204, §§201–206 (US).

[32] Akbulut C, et al. *Evaluating Language Models for Harmful Manipulation*. arXiv:2603.25326, 2026.

[33] Phuong M, et al. *Evaluating Frontier Models for Stealth and Situational Awareness*. arXiv:2505.01420, 2025.

[34] Rodriguez M, et al. *A Framework for Evaluating Emerging Cyberattack Capabilities of AI*. arXiv:2503.11917, 2025.

[35] Balesni M, et al. *Towards Evaluations-Based Safety Cases for AI Scheming*. arXiv:2411.03336, 2024.

[36] Barnett P, Thiergart L. *What AI Evaluations for Preventing Catastrophic Risks Can and Cannot Do*. arXiv:2412.08653, 2024.

[37] Korbak T, et al. *Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety*. arXiv:2507.11473, 2025.

[38] Anthropic. *Natural Language Autoencoders Produce Unsupervised Explanations of LLM Activations*. Transformer Circuits, May 2026. Available at: https://transformer-circuits.pub/2026/nla/index.html

[39] UK AI Security Institute. *What can sandboxed AI agents learn about their evaluation environments?* AISI, 20 April 2026. Available at: https://www.aisi.gov.uk/blog/what-can-sandboxed-ai-agents-learn-about-their-evaluation-environments

[40] International Network of AI Safety Institutes. *International Joint Testing Exercise: Agentic Testing*. 17 July 2025. Available at: https://www.aisi.gov.uk/blog/international-joint-testing-exercise-agentic-testing

[41] NIST Center for AI Standards and Innovation (CAISI). *NIST AI 800-2: Practices for Automated Benchmark Evaluations of Language Models* (Initial Public Draft). 30 January 2026.

[42] California State Legislature. *Transparency in Frontier Artificial Intelligence Act* (SB 53). Effective 1 January 2026.

[43] US House of Representatives. *Great American Artificial Intelligence Act of 2026* (GAAIA), discussion draft, 4 June 2026 (Reps. Obernolte and Trahan).

[44] *Executive Order 14365* (US), "Ensuring a National Policy Framework for Artificial Intelligence," 11 December 2025, directing federal challenges to state AI laws and establishing an AI Litigation Task Force under the Attorney General.

[45] *New Delhi Declaration on AI Impact*, India AI Impact Summit, February 2026.

[46] METR. *Frontier Risk Report (February–March 2026)*. METR, 19 May 2026. Available at: https://metr.org/blog/2026-05-19-frontier-risk-report/

[47] Feakins S, Habli I, Morgan P. *Clear, Compelling Arguments: Rethinking the Foundations of Frontier AI Safety Cases*. arXiv:2603.08760, 2026.

[48] New York State. *RAISE Act* (S6953B/A6453B). Signed 19 December 2025; effective 1 January 2027.

[49] Dung L, Mai F. *AI Alignment Strategies from a Risk Perspective: Independent Safety Mechanisms or Shared Failures?* arXiv:2510.11235, 2025.

[50] Bengio Y, et al. *International AI Safety Report 2026*. arXiv:2602.21012, 2026.

[51] Srivastava V. *Fundamental Limits of Black-Box Safety Evaluation*. arXiv:2602.16984, 2026.

[52] Shah R, Flynn F, et al. *Securing the Future of AI Agents* (AI Control Roadmap). Google DeepMind, 18 June 2026. Available at: https://deepmind.google/blog/securing-the-future-of-ai-agents/

[53] Davidsen Buhl M, Pfau J, Hilton B, Irving G. *An Alignment Safety Case Sketch Based on Debate*. arXiv:2505.03989, 2025.

[54] Vishwarupe V, Shadbolt N, Jirotka M, Flechais I. *The Evaluation Differential: When Frontier AI Models Recognise They Are Being Tested*. arXiv:2605.11496, 2026.

[55] Knight JC, Leveson NG. *An Experimental Evaluation of the Assumption of Independence in Multiversion Programming*. IEEE Transactions on Software Engineering, SE-12(1):96–109, January 1986.

[56] Ron J, Baudry B, Monperrus M. *N-Version Programming with Coding Agents*. arXiv:2606.20158, 2026.

[57] Shlegeris B. *Efficient Tradeoffs and the Safety-Usefulness Tradeoff Model*. Redwood Research, 8 June 2026. Available at: https://blog.redwoodresearch.org/p/efficient-tradeoffs-and-the-safety

[58] Stelling et al. (SaferAI). *Evaluating AI Providers' Frontier Safety Frameworks*. arXiv:2512.01166, 2025.

[59] METR. *Common Elements of Frontier AI Safety Policies*. METR, 9 December 2025. Available at: https://metr.org/blog/2025-12-09-common-elements-of-frontier-ai-safety-policies/

[60] Anthropic. *Transparency Hub: Model Report* (last updated 20 February 2026). Available at: https://www.anthropic.com/transparency/model-report

[61] Hamin M, Edelman B (NIST Center for AI Standards and Innovation). *Cheating on AI Agent Evaluations*. CAISI Research Blog, 2 December 2025. Available at: https://www.nist.gov/blogs/caisi-research-blog/cheating-ai-agent-evaluations

[62] California State Legislature. *SB 53: Transparency in Frontier Artificial Intelligence Act*. 2025. Available at: https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202520260SB53

[63] European Union. *Artificial Intelligence Act (Regulation 2024/1689), Article 73: Reporting of Serious Incidents*. Official Journal, 12 July 2024. Available at: https://artificialintelligenceact.eu/article/73/

[64] Anthropic. *Claude Fable 5 and Claude Mythos 5*. 9 June 2026, with service notices of 12 June 2026 (suspension of access) and 1 July 2026 (restoration). Available at: https://www.anthropic.com/news/claude-fable-5-mythos-5

[65] The Hacker News. *Anthropic Restores Claude Fable 5 After U.S. Lifts Jailbreak-Linked Export Controls*. 1 July 2026. Available at: https://thehackernews.com/2026/07/anthropic-restores-claude-fable-5-after.html

AI-assistance disclosure. The authors used AI assistance in preparing this paper, including literature retrieval, reference verification, and drafting iteration. The arguments and their defects remain the authors' responsibility.