What Changes Faster Than It Can Be Checked: Reading Frontier AI Safety Frameworks as Control Loops

An organisational-control method for analysing published frontier safety frameworks, with three findings from primary documents

Maziyar Bahrami
PhD Researcher in Management, Luiss Guido Carli University
August 2026

Abstract

Frontier AI developers now publish safety frameworks combining capability thresholds, evaluations, safeguards, decision procedures, monitoring, and assurance commitments. Existing analyses evaluate these documents stage by stage, asking whether each risk-management function is present and adequately specified. This essay argues for a complementary reading in which the framework is treated as an organisational control loop and the objects of analysis are the dependencies between stages and the rate at which the framework itself changes. The stage model used here is not novel and is described as such. What the reading produces is three findings from primary documents: that Anthropic's external-review trigger depends on an internal-distribution rule which a later version narrowed without either change being presented as affecting the other; that OpenAI maintains an operational framework and a statute-facing framework in parallel, so that external coding of "the framework" may address the wrong artefact; and that Anthropic's policy moved through five versions between February and July 2026 while its commissioned compliance review runs approximately annually, so that the object of assurance now changes faster than the assurance cycle. The essay states what evidence would show the method to be redundant, and sets out research questions answerable from public documents.

Keywords: frontier AI governance; frontier AI safety frameworks; organisational control; AI assurance; frontier AI auditing; governance latency

Source note. All academic references and developer documents cited here were checked against arXiv, publisher, or official developer sources in August 2026. Developer framework versions are as of 14 August 2026. Where a claim rests on a document's own text rather than on independent evidence of practice, the text says so.

1. A policy that moves through five versions in nineteen weeks

Anthropic's Responsible Scaling Policy reached version 3.0 on 24 February 2026. It then moved through version 3.1 on 2 April, 3.2 on 29 April, 3.3 on 26 May, and 3.4 on 8 July. Each change is logged, each is accompanied by a published redline, and each has a stated rationale (Anthropic, 2026c). By the standards of corporate risk documentation this is unusually transparent versioning practice.

The same policy commits the company to commission a third-party review, on approximately an annual basis, of whether it adhered to the policy's main procedural commitments. That review is explicitly limited: it addresses *procedural compliance, not substantive outcomes* (Anthropic, 2026a, §4).

Set those two facts beside each other and a governance question appears that neither one raises alone. An annual review of procedural compliance with a policy that moved through five versions during the

period must decide which version's procedures it is assessing. Anthropic's public documents do not say. The point is not that the company has done something wrong; version-scoping may well be handled in the engagement letter. The point is that this is a real control question, it is generated by the interaction between two commitments rather than by either, and no published analysis of frontier safety frameworks asks it, because published analyses evaluate frameworks one function at a time.

That is the argument of this essay. Frontier safety frameworks are now mature enough that the interesting properties are relational rather than compositional. Whether a framework contains a risk-identification function is largely settled and has been well measured. Whether the rule governing one function quietly determines how reliably another function operates is not measured at all, and is visible in the documents that already exist.

I set out the method, state plainly what it borrows, apply it to three developers, and say what would show it to be redundant.

2. What the literature has already established

The field is not short of structure. Buhl, Bucknall and Masterson (2025) catalogue recurring practices across frontier safety frameworks. Campos et al. (2025) map frontier AI onto established risk-management practice and propose key risk and key control indicators to connect abstract tolerances to observable evidence. Ziosi et al. (2026) survey open problems across the risk-management process and make the sharper observation that emerging frontier safety practices are sometimes misaligned with, and can cut against, established risk-management frameworks rather than simply lagging them.

The measurement work is further along than the theory. Stelling et al. (2026) assess twelve published frameworks against sixty-five weighted criteria across risk identification, analysis and evaluation, treatment, and governance. Overall scores run from 34 per cent for Anthropic to 8 per cent for Cohere, with a median of 18 per cent, and a hypothetical company adopting every leading practice already present somewhere in the peer set would reach 54 per cent. Their central conclusion is that commitments are frequently vague enough that an outsider cannot predict company decisions, judge whether planned responses are adequate, or tell whether a commitment has been kept.

Assurance research asks who checks. Homewood et al. (2025) analyse third-party compliance review and the access problem that makes it hard. Brundage et al. (2026) define frontier AI auditing as third-party verification based on deep, secure access to non-public information, argue that audits should cover organisation-level matters including internal deployment, information security, and safety decision-making processes, and propose four AI Assurance Levels running from a time-bounded system audit up to continuous verification designed to withstand deliberate deception. Schuett (2025) argues for an internal audit function with the organisational independence the third line of the Institute of Internal Auditors' model presupposes, and makes the incentive argument directly: because a pause is expensive, the people judging whether safeguards are adequate have reasons to conclude that they are, and superficial compliance is a live risk.

Systems safety supplies the control vocabulary. Leveson (2012) treats safety as a control problem in sociotechnical systems rather than as component failure. Mylius (2025) applies Systems-Theoretic Process Analysis to frontier AI, deriving unsafe control actions and loss scenarios for a published AI control safety case, and observes that frontier companies do not describe structured hazard-identification methods in any detail.

The International AI Safety Report 2026 supplies the environmental conditions. It records that models increasingly distinguish test settings from deployment and exploit loopholes in evaluations, so that dangerous capabilities can go undetected before release. It separately identifies an evaluation gap running the other way, in which benchmark performance overstates real-world utility. It notes that competitive pressure can push developers to reduce investment in testing and mitigation in order to ship faster, that

commercial incentives keep the most informative material proprietary, and that governance institutions adapt slowly because technical capacity takes time to build (Bengio et al., 2026).

Taken together this is not a field lacking mechanisms. It is a field in which mechanisms have accumulated faster than any account of how they interact.

3. Why a feedback loop is not a contribution

It would be easy, and wrong, to propose that frontier AI governance needs a feedback loop.

NIST's AI Risk Management Framework already describes Govern, Map, Measure and Manage as interconnected functions operating continuously across the lifecycle (Tabassi, 2023). Kleiner (2026) already models AI governance as two coupled feedback loops, an outer loop of societal oversight and an inner loop of operational deployment, and derives design principles of distributed authority, requisite variety and reflexive checkpoints from cybernetics and polycentric governance. Those principles overlap substantially with anything one might propose about authority, adaptability and verification. Muhammad, Yow and Alsenan (2026) translate governance requirements into versioned policy specifications and deterministic gates in continuous integration pipelines. Lupo, Vo and Locke (2026) argue that point-in-time, document-based audit cannot keep pace with fast-changing agentic systems, which is the latency problem stated in assurance terms.

So the six-function cycle used in this essay — sense, interpret, decide, act, verify, learn — is not offered as a discovery. It is a relabeling of standard control-cycle stages, and every stage corresponds to activity already named in NIST, in frontier risk-management proposals, in safety-case research, and in audit scholarship. Its only role here is to give the transitions names so that dependencies between them can be discussed.

The contribution has to be somewhere else. It is in two places.

First, in the unit of analysis. Existing evaluation of frontier frameworks is *stage-wise*: it scores whether risk identification is adequate, whether thresholds are specified, whether governance responsibilities are allocated. This essay reads *cross-stage*: it asks which rules governing one function determine the reliability of another. That is a different operation on the same documents, and Section 6 shows it returns different results.

Second, in treating the framework's own revision rate as a governed variable rather than as evidence of responsiveness. Frameworks are now revised several times a year. Revision is normally reported as a virtue, and often is. But a document that changes faster than the mechanisms designed to check it creates a coordination problem that neither the document nor the checking mechanism is written to handle.

Where this differs from Kleiner is scale and object. Reflexive Polycentric Control couples a societal regulatory loop to an operational one and asks how regime type conditions both. The reading here stays inside a single developer, takes the published framework as the artefact, and asks about dependencies among internal functions. Where it differs from Brundage et al. is access. They ask what an auditor could establish with deep non-public access; this asks what an outside researcher can establish from documents that are already public. Those are complementary, and the second is the only one available to almost everyone.

4. Traceability is the precondition, not one condition among many

Before any dependency can be examined, something has to be reconstructable. Traceability means that a reviewer can follow a chain from signal to interpretation, decision, action, and verification, with versions, assumptions, approvals, dissent, and subsequent changes preserved.

This is not documentation for its own sake, and it is not one property in a list. It is the condition under which the other properties become assessable at all. Anthropic's practice makes the point concretely. Because the company publishes redlines and a change log for every RSP version, an outside reader can see exactly which sentence changed in July 2026 and what the company said its reason was. That is what makes the analysis in Section 6 possible. Where a developer publishes only a current version, the same analysis cannot be performed, and the absence of a finding would say nothing about the underlying governance.

Traceability therefore does double duty. It is a governance property worth having, and it is the methodological precondition for external research on governance properties. Researchers should expect their ability to study a developer to correlate with that developer's document hygiene rather than with its actual control quality, which is a serious selection problem and is discussed in Section 7.

5. Four conditions on the transitions

Given traceability, four properties determine whether a transition between functions carries load.

Authority. A signal that reaches someone without decision rights has not reached the loop. The question is not whether a safety function exists but what it can stop. Published frameworks vary in how far they answer this. Anthropic's RSP is unusually specific: internal experts draft a Risk Report, separate internal reviewers critique it, and the CEO and Responsible Scaling Officer give final approval and make the determination on adequacy and on downstream development or deployment plans, after which the Board and Long-Term Benefit Trust are notified. Board and Trust *approval* becomes mandatory only in one circumstance — when the decision to proceed rests substantially on a marginal-risk argument, meaning an argument that the company's own systems add relatively little to risk already imposed by others (Anthropic, 2026a, §3.4).

That is a precisely drawn authority rule, and its structure is worth noticing. The trigger for escalating approval above executive level is not the *severity* of the assessed risk but the *type of argument* used to justify proceeding. A determination that risk is low escalates no further than the CEO and RSO regardless of what is at stake; a determination that risk is high but comparatively marginal must go to the Board and the Trust. There may be good reasons for that design: marginal-risk reasoning is exactly the reasoning most vulnerable to motivated construction, so subjecting it to additional scrutiny is defensible. But the rule means that severity alone never moves the decision, and no stage-wise rubric asking "are governance responsibilities allocated?" would surface this, because the responsibilities are allocated, clearly and in writing.

Green and Welsh (1988) supply the reason this matters theoretically. Their reframing of control through cybernetics insists on dependence relations among actors, not only on information flow. Receiving a signal and being positioned to act on it are different conditions, and the second is a function of organisational dependence rather than of information design.

Independence. Assurance requires distance from the thing assured. Frontier AI makes this harder than usual, because meaningful assessment requires deep technical access, and deep access tends to come with proximity, dependence, and shared assumptions. Brundage et al. (2026) build their assurance ladder around exactly this trade-off: higher assurance levels are defined by progressively deeper access, up to continuous verification designed to detect active deception.

The practical question is therefore not how to maximise separation but how much epistemic independence survives the access relationship. Anthropic's RSP addresses this directly in its criteria for selecting external reviewers, requiring that reviewers not be organisations whose revenue and standing depend entirely on continued work with companies like Anthropic, and excluding financial interests and close personal ties. Version 3.2, effective 29 April 2026, went further by authorising the Long-Term Benefit Trust to request external review and to approve the choice of reviewer. Moving reviewer selection

away from the reviewed party is a textbook independence improvement, and it is worth recording that framework revision in this period moved this condition in the strengthening direction.

Latency. How long the loop takes, relative to how fast the thing being controlled changes. This is the condition the literature names least and needs most. Ashby's (1956) requisite variety argument gives the general form: a controller must be able to generate responses at the rate and range the disturbance requires. Kleiner (2026) carries the same principle into AI governance. The International AI Safety Report 2026 supplies the empirical case that the rates are mismatched, noting both that competitive pressure compresses testing and that governance institutions adapt slowly (Bengio et al., 2026).

Latency has an underappreciated second-order form, which Section 6 develops: not the time from signal to action, but the time from framework change to assurance of the changed framework.

Revisability. Frameworks must change as evidence changes, and revision is where the sharpest tension in the whole system sits. A framework that cannot be revised becomes obsolete. A framework that can be revised freely can be revised precisely when a commitment becomes expensive.

Public evidence shows revision running in both directions. DeepMind's Frontier Safety Framework added a Critical Capability Level for harmful manipulation in its September 2025 revision, extended safety-case review to large-scale internal deployments for machine-learning R&D capability levels, and in version 3.1 of 17 April 2026 introduced Tracked Capability Levels, a deliberately earlier detection tier beneath the critical thresholds (Google DeepMind, 2026). OpenAI's Preparedness Framework version 2 of April 2025 moved in the other direction on scope, retiring the persuasion category and the Low and Medium threshold tiers. Anthropic's version 3.0 restructured the policy away from unconditional mitigation commitments toward a three-part scheme separating company plans, industry-wide recommendations, and commitments contingent on competitor behaviour, on the stated reasoning that unilateral pausing could leave the least cautious developers setting the pace.

The interesting question is not whether a given change loosens or tightens. It is whether the change is legible enough that the question can be asked. Anthropic's own change log offers the best available illustration of both sides. In October 2024 the company published four instances in which it had fallen short of the prior policy's requirements, one of which was that a scheduled evaluation ran three days past a three-month interval; the resolution was to extend the interval to six months, on the reasoning that a three-month cadence had been pushing teams toward frequent rather than thorough evaluation. Read one way this is a compliance gap closed by relaxing the requirement. Read another way it is exactly the learning a control loop is supposed to produce, since a cadence that degrades evaluation quality is a badly specified control. Both readings are available *only because the shortfall and the reasoning were published*. Where they are not published, neither reading is possible, and silence is indistinguishable from either.

Anthropic's February 2026 determination that Claude Opus 4.6 did not cross its automated-R&D threshold makes the same point about interpretation rather than revision. The company recorded that ruling the threshold out confidently was becoming harder and required judgments more subjective than it would like, and responded by committing to additional sabotage risk reporting for subsequent frontier models. Threshold governance depends on a determination that a threshold has not been crossed, and that determination is itself an act of judgment under acknowledged uncertainty.

I have not included effectiveness as a separate condition, because it is not a property of transitions. It is a property of what verification tests, and it belongs in Section 6.

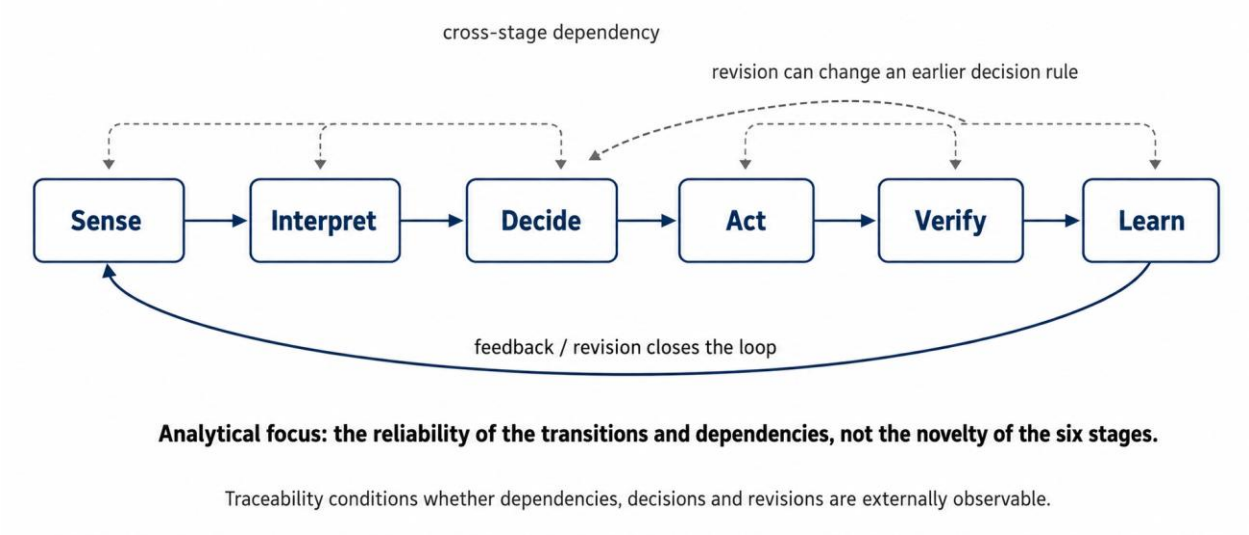


Figure 1. Cross-stage dependencies in a governance control loop. The six functions are standard; the analytical objects are the transitions between them and cross-stage dependencies. Traceability conditions what is externally observable.

6. Three findings from public documents

What follows are readings of primary documents. They establish what the documents say and how their provisions interact. They establish nothing about whether the arrangements work, because published policy is evidence of stated intent and is not evidence of practice.

Anthropic: a verification trigger resting on a sensing rule.

Version 3.0 of the RSP makes full external review of a Risk Report mandatory when two conditions hold together: the report covers a model the company judges highly capable, and the report is significantly redacted, meaning redactions omit material a reasonable external safety researcher would need to assess the overall risk. Whether a report is significantly redacted is determined by the RSO, CEO, Board, or Trust. The policy then adds a supporting observation: unredacted reports go to a large number of employees, who are positioned to raise concerns if the public version fails that standard (Anthropic, 2026a, §3.6).

That supporting observation is load-bearing. It supplies a distributed check on a judgment that would otherwise be made entirely by the parties with the strongest interest in the answer. In control terms, the reliability of a verification trigger is partly underwritten by a sensing population.

Version 3.4, effective 8 July 2026, changed the internal transparency provision. Where version 3.0 committed to sharing final unredacted Risk Reports with regular-clearance staff, version 3.4 removed that requirement and substituted a floor of at least two hundred employees (Anthropic, 2026b).

Whether this matters is genuinely unclear, and I want to be careful. Two hundred people is a substantial and diverse readership, quite possibly ample for the purpose. The change was published, logged, and redlined, which is more than most organisations do. And there are obvious legitimate reasons to narrow distribution of unredacted catastrophic-risk material as a company grows.

The finding is not that a commitment was weakened. It is that a rule in the internal transparency section conditions the reliability of a trigger in the external review section, that the two are separated by several pages and never cross-referenced, and that the change was described as a change to internal sharing rather than as a change to the external review mechanism. A rubric asking whether the framework specifies external review would score this framework identically before and after. A rubric asking whether internal transparency is specified would register a narrowing but would not connect it to anything. Only reading the document as a loop connects them.

The same document contains a second cross-stage dependency worth flagging. Appendix A sets out commitments conditional on competitor behaviour, several of which are triggered by states of the world such as having clear evidence that no competitor will soon develop a comparably capable model. The appendix states candidly that the company will often lack enough information to determine whether the relevant scenario applies. A decision rule whose trigger condition the organisation acknowledges it frequently cannot observe is a decision rule with an unresolved sensing dependency, and this is specific to frontier AI in a way that ordinary risk thresholds are not, because the trigger is a fact about a competitor's internal state rather than about the organisation's own system.

OpenAI: two documents, two functions, one name.

OpenAI published a Frontier Governance Framework on 28 May 2026. It is easy to read this as a successor or expansion of the Preparedness Framework. It is neither. OpenAI describes it as a public document explaining how existing safety and security practice aligns with emerging legal requirements, specifically California's Transparency in Frontier AI Act and the EU AI Act's Code of Practice for General-Purpose AI, and states that the Preparedness Framework remains the foundation for how the company defines and operates its approach to the most serious risks, including internal practices beyond current legal requirements (OpenAI, 2026).

So two artefacts now coexist. One is the operational control document. The other is the statute-facing disclosure document, structured around the categories a regulator asks for rather than around the company's own decision process.

This has direct methodological consequences for the empirical literature. Stelling et al. (2026) score published frameworks; the Frontier AI Safety Commitments agreed at Seoul in 2024 anticipate that published frameworks carry accountability weight (DSIT, 2024); emerging legislation treats them as external accountability instruments. If the artefact that satisfies the statutory definition is not the artefact that governs internal decisions, then coding "the framework" is coding a translation, and differences between developers may partly reflect differences in which layer they have chosen to publish rather than differences in governance. This is not a criticism of OpenAI, which has been explicit about the relationship. It is a warning about inference, and it applies to every study that treats published frameworks as comparable objects.

Cadence: the object of assurance changing faster than the assurance cycle.

Return to the observation in Section 1. Anthropic's RSP moved through five versions between 24 February and 8 July 2026, a span of nineteen weeks. The policy commits to an approximately annual commissioned third-party review of adherence to its main procedural commitments, expressly limited to procedural compliance rather than substantive outcomes (Anthropic, 2026a, §4).

Three things follow, and they are separable.

The first is a scoping problem. A review conducted annually against a policy revised five times must select a version, a date range, or some rule for handling requirements that existed for part of the period. None of these is stated publicly. This is soluble and may already be solved internally; the point is that it is not visible, and it is the kind of thing that determines whether an assurance finding means anything.

The second is a coverage problem specific to the compliance/effectiveness boundary. Anthropic's own text draws that boundary explicitly. A procedural review asks whether the company did what the framework required. It does not ask whether the resulting safeguards reduce the risks they target. When the framework changes frequently, procedural review becomes progressively less informative about the underlying risk position, because a rising share of what it checks is compliance with recently written requirements whose substantive adequacy has not been assessed by anyone.

The third is the general form. Ashby's requisite variety argument says a controller must match the variety of what it controls. Here the controlled process includes the framework itself. Assurance operating on an annual cycle over an artefact revised quarterly or faster is under-varied with respect to its object.

Lupo, Vo and Locke (2026) make the parallel argument for agentic systems, that point-in-time document-based audit cannot scale to rapidly changing control environments. The frontier framework case is the same structure applied to the governance document rather than to the system.

I have stated this for one developer because Anthropic publishes the version history that makes it computable. Whether comparable mismatches exist across the other eleven published frameworks is an open empirical question and one of the most tractable in Section 9.

7. What would show this reading to be redundant

A diagnostic that cannot fail is not a diagnostic. Four things would substantially undercut the argument.

If experienced analysts applying a stage-wise rubric independently identified the same cross-stage dependencies, then the loop reading contributes nothing beyond vocabulary. This is directly testable: give two groups the same framework, one a stage-wise instrument and one a dependency instrument, and compare what each surfaces.

If compliance reviews already scope explicitly by policy version, the cadence finding dissolves into a solved problem. A single disclosure would settle it.

If cross-stage dependencies turn out not to predict anything about observed governance outcomes — if frameworks with tightly coupled provisions behave no differently from frameworks with loosely coupled ones — then the coupling is a textual property without organisational significance.

And if the dependencies identified here are artefacts of document drafting rather than of organisational design, the method is reading the wrong object. This is the most serious risk. A framework is a text produced by people with limited time. Two provisions may fail to cross-reference each other because nobody noticed, not because the organisation has an unmanaged dependency. Distinguishing these requires evidence the documents cannot supply.

There is also a selection problem worth stating plainly. The method works best on developers who publish redlines and change logs. Anthropic yields the most findings here precisely because it discloses the most, and it also scores highest in Stelling et al.'s assessment. Nothing in this essay supports any conclusion about which developer manages risk better. If anything, the analysis punishes disclosure, and readers should resist the inference that the developer generating the most findings is the one with the most problems.

8. What this changes for practice

For developers, the operational suggestion is narrow. Take the framework's own text and build a dependency map: for each provision that conditions a decision, identify which other provision supplies the information the condition requires. Anthropic's external review trigger depending on internal distribution is one instance; competitor-contingent commitments depending on unobservable competitor states is another. The exercise takes a day and needs no new data.

For internal audit and assurance functions, the compliance/effectiveness boundary is already drawn explicitly in at least one published framework, which makes it a defensible scoping question rather than a critique. The additional question raised here is versioning: what policy version is the engagement testing against, and how are requirements that changed mid-period handled.

For third-party reviewers, the cadence mismatch suggests that annual engagement may be the wrong shape for an artefact revised quarterly. Brundage et al.'s (2026) assurance ladder already anticipates continuous forms at higher levels; the case for continuity here comes not from model capability but from document volatility, which is a cheaper and more immediate argument.

For regulators and AI safety and security institutes, the OpenAI case raises a specific question. If statutory regimes require publication of a frontier framework, developers have a reasonable incentive to produce a document optimised for that requirement. Whether the statute-facing document and the operative document are the same is a question a regulator can ask directly and an outside researcher cannot.

9. Research questions

The following are answerable, and the first three need only public documents.

Do framework revision cadences exceed assurance cadences across the twelve published frontier safety frameworks? Version histories and stated review commitments are largely public. The measure is simple: revisions per year against committed review frequency. This would establish whether the Anthropic case is characteristic or exceptional.

How many cross-stage dependencies can be identified in published frameworks, and do they cluster? Coding for provisions whose operation depends on information supplied by another provision would produce the first dependency map of this literature. The prediction implied here is that dependencies cluster around verification triggers, because triggers require a judgment and judgments require inputs.

Where developers maintain both a statute-facing and an operational framework, how far do they diverge? OpenAI's two documents are both public. A provision-level comparison would establish whether regulatory translation preserves or reshapes the governance content, and would calibrate how much confidence the empirical literature should place in published frameworks as a data source.

Does formal specification of decision rights predict escalation behaviour? This needs interviews or process records. Anthropic's rule that argument type rather than risk severity determines board-level approval is a specific, testable design; whether such rules bind under pressure is the central unanswered question of the whole area.

What does external review of a safety case actually produce? Barrett et al. (2026) provide one of the first published detailed accounts of external review of a frontier-AI safety case, applying an Assurance 2.0-based review to DeepMind's scheming-inability case. Replication across cases would move safety-case assurance from proposal to evidence.

10. Limits

The method reads texts. It cannot establish that any organisation behaves as its documents describe, and it cannot rank developers on safety. Its findings are conditional on disclosure, which biases it toward developers who disclose. Its central case rests on one company's version history. The stage model it uses is borrowed and adds nothing on its own.

What it does is change which questions get asked of documents that are already public and already being coded. Existing instruments ask whether the parts are present and adequately specified, and answer that most frameworks are underspecified. The reading here asks how the parts depend on one another and how fast the whole thing moves, and finds at least three things worth knowing that the stage-wise question does not reach. A trigger can be well specified and still rest on a rule that changed elsewhere. A framework can be complete and still be the wrong document. A revision can be an improvement and still outrun the mechanism built to check it.

None of that is settled by reading. It is a reason to look.

Table 1. Two ways of reading the same framework provision

Provision	Stage-wise question	Cross-stage question

Provision	Stage-wise question	Cross-stage question
External review triggered when a report is "significantly redacted"	Is external review specified, and by whom is the judgment made?	What supplies the information that makes the judgment reliable, and has that provision changed?
Commitments contingent on competitor behaviour	Are the commitments specified with clear triggers?	Can the organisation observe the trigger condition, and what happens when it cannot?
Annual third-party procedural compliance review	Is independent review committed to, and how often?	How often does the reviewed policy change, and which version is reviewed?
Capability threshold with required mitigations	Is the threshold defined and the mitigation specified?	Who determines that the threshold has not been crossed, and what does that determination depend on?

Suggested citation: *Bahrami, M. (2026). What changes faster than it can be checked: Reading frontier AI safety frameworks as control loops. Research essay.*

References

Anthropic. (2026a). Responsible Scaling Policy (Version 3.0, effective February 24, 2026). https://www.anthropic.com/responsible-scaling-policy/rsp-v3-0

Anthropic. (2026b). Responsible Scaling Policy (Version 3.4, effective July 8, 2026). https://cdn.sanity.io/files/4zrzovbb/website/0bacdc8440ea96e62a8766d99ebe1d4eea6d5f3a.pdf

Anthropic. (2026c). Responsible Scaling Policy: Current and prior versions and update log (through Version 3.4, effective July 8, 2026). https://www.anthropic.com/responsible-scaling-policy

Ashby, W. R. (1956). An introduction to cybernetics. Chapman & Hall.

Barrett, S., Campos Zabala, F. J., Fillingham, S. P., Siddique, U., Walpole, J., Bloomfield, R., & Papadatos, H. (2026). Lessons from external review of DeepMind's scheming inability safety case. arXiv:2604.21964. https://arxiv.org/abs/2604.21964

Bengio, Y., Clare, S., Prunkl, C., Murray, M., Andriushchenko, M., Bucknall, B., Bommasani, R., Casper, S., Davidson, T., Douglas, R., Duvenaud, D., Fox, P., Gohar, U., Hadshar, R., Ho, A., Hu, T., Jones, C., Kapoor, S., Kasirzadeh, A., et al. (2026). International AI Safety Report 2026. https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026

Brundage, M., Dreksler, N., Homewood, A., McGregor, S., Paskov, P., Stosz, C., Sastry, G., Cooper, A. F., Balston, G., Adler, S., Casper, S., Anderljung, M., et al. (2026). Frontier AI auditing: Toward rigorous third-party assessment of safety and security practices at leading AI companies. arXiv:2601.11699. https://arxiv.org/abs/2601.11699

Buhl, M. D., Bucknall, B., & Masterson, T. (2025). Emerging practices in frontier AI safety frameworks. arXiv:2503.04746. https://arxiv.org/abs/2503.04746

Campos, S., Papadatos, H., Roger, F., Touzet, C., Quarks, O., & Murray, M. (2025). A frontier AI risk management framework: Bridging the gap between current AI practices and established risk management. arXiv:2502.06656. https://arxiv.org/abs/2502.06656

Department for Science, Innovation and Technology. (2024, May 21). Frontier AI Safety Commitments, AI Seoul Summit 2024. GOV.UK. https://www.gov.uk/government/publications/frontier-ai-safety-commitments-ai-seoul-summit-2024/frontier-ai-safety-commitments-ai-seoul-summit-2024

Google DeepMind. (2026). Frontier Safety Framework (Version 3.1, April 17, 2026). https://deepmind.google/frontier-safety/

Green, S. G., & Welsh, M. A. (1988). Cybernetics and dependence: Reframing the control concept. Academy of Management Review, 13(2), 287–301. https://doi.org/10.5465/amr.1988.4306891

Homewood, A., Williams, S., Dreksler, N., Lidiard, J., Murray, M., Heim, L., Ziosi, M., Ó hÉigeartaigh, S., Chen, M., Wei, K., Winter, C., Brundage, M., Garfinkel, B., & Schuett, J. (2025). Third-party compliance reviews for frontier AI safety frameworks. arXiv:2505.01643. https://arxiv.org/abs/2505.01643

Kleiner, J. (2026). Governance of AI and by AI: Feedback loops, regime variation and reflexive polycentric control. Global Policy. https://doi.org/10.1111/1758-5899.70177

Leveson, N. G. (2012). Engineering a safer world: Systems thinking applied to safety. MIT Press.

Lupo, G., Vo, B. Q., & Locke, N. (2026). Trustworthy AI Posture (TAIP): A framework for continuous AI assurance of agentic systems at horizontal and vertical scale. arXiv:2603.03340. https://arxiv.org/abs/2603.03340

Muhammad, A. E., Yow, K.-C., & Alsenan, S. (2026). Audit-as-code: A policy-as-code framework for continuous AI assurance. Frontiers in Artificial Intelligence, 9, 1759211. https://doi.org/10.3389/frai.2026.1759211

Mylius, S. (2025). Systematic hazard analysis for frontier AI using STPA. arXiv:2506.01782. https://arxiv.org/abs/2506.01782

OpenAI. (2025, April 15). Preparedness Framework (Version 2). https://openai.com/index/updating-our-preparedness-framework/

OpenAI. (2026, May 28). OpenAI's Frontier Governance Framework. https://openai.com/index/openai-frontier-governance-framework/

Schuett, J. (2025). Frontier AI developers need an internal audit function. Risk Analysis, 45(6), 1332–1352. https://doi.org/10.1111/risa.17665

Stelling, L., Murray, M., Galizzi, B., Schaffelder, M., Campos, S., & Papadatos, H. (2026). Evaluating AI providers' frontier safety frameworks (Version 5). arXiv:2512.01166. https://arxiv.org/abs/2512.01166v5

Tabassi, E. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0) (NIST AI 100-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1

Ziosi, M., Plueckebaum, M., Casper, S., Papadatos, H., Chin, Z. S., Slattery, P., Gealy, J., Rudner, T. G. J., Tse, B., Gil, A., Paskov, P., et al. (2026). Open problems in frontier AI risk management. arXiv:2604.25982. https://arxiv.org/abs/2604.25982