
Self-Governance Under Revision: How Frontier AI Safety Commitments Change¹

Ajay Agarwal
University of Fukui

Daniel Samuel Polak
Independent Researcher

With
Apart Research

Abstract

Frontier AI labs have published voluntary safety frameworks committing to evaluate dangerous capabilities and implement safeguards before deployment. These documents, including Anthropic's Responsible Scaling Policy, OpenAI's Preparedness Framework, and Google DeepMind's Frontier Safety Framework, are often cited as evidence of responsible self-governance, yet can be revised at any time without external approval.

We developed a taxonomy of commitment changes and applied it to over 60 revisions across three major labs, coding direction (strengthening, weakening, neutral), mechanism, and changelog disclosure. We find sharp divergence: OpenAI weakened 19 commitments with zero strengthenings; Anthropic was roughly balanced (11 weakenings, 10 strengthenings); and DeepMind net strengthened (4 weakenings, 11 strengthenings). Strengthenings were twice as likely to be omitted from changelogs as weakenings (68% vs. 38%), and all three labs weakened evaluation-frequency commitments. Most weakenings reduced oversight-relevant properties: 68% affected external accountability and 56% reduced measurability.

Overall, framework evolution varies substantially by lab, while official communications systematically underreport commitment reductions. Our taxonomy and dataset enable ongoing monitoring

¹ Research conducted at the [The Technical AI Governance Challenge | Apart Research](#), 2026

and inform policymakers on which commitments may require regulatory protection rather than voluntary maintenance.

Keywords: AI governance, safety frameworks, responsible scaling, frontier AI, commitment tracking, regulatory policy

1. Introduction

Frontier AI labs have published safety frameworks committing to evaluate dangerous capabilities and implement safeguards before deployment. These voluntary commitments, including Anthropic's Responsible Scaling Policy, OpenAI's Preparedness Framework, and DeepMind's Frontier Safety Framework, are increasingly referenced by policymakers as evidence of responsible self-governance ([METR, 2025b](#)).

However, these frameworks have no built-in or external mechanisms preventing modification. Labs can update them at any time without approval. This flexibility enables necessary adaptation as capabilities and understanding evolve but also allows commitments to be quietly weakened, removed, or made vaguer. We ask: In which direction has this flexibility been exercised?

To answer this, we developed a systematic taxonomy of how safety commitments change and applied it to over 60 documented revisions across three major labs. We tracked each change's direction (strengthening, weakening, or neutral), mechanism (how it changed), and whether it was disclosed in official changelogs.

Our analysis reveals significant divergence across labs, asymmetric transparency about weakenings versus strengthenings, and systematic vulnerability in specific commitment types findings with direct implications for AI governance policy.

2. Methods

Scope and Document Selection

We analyzed the complete version histories of three major frontier AI safety frameworks: Anthropic's Responsible Scaling Policy (versions 1.0, 2.0, 2.1, and 2.2, spanning September 2023 to May 2025), Google DeepMind's Frontier Safety Framework (versions 1.0, 2.0, and 3.0, spanning May 2024 to May 2025), and OpenAI's Preparedness Framework (versions 1.0 and 2.0, spanning December 2023 to April 2025). These frameworks were selected because they represent the three leading frontier AI developers, have published multiple versions enabling longitudinal analysis, and explicitly address catastrophic risk governance.

Unit of Analysis

The unit of analysis is a discrete commitment change: a modification to a specific obligation, threshold, requirement, or governance process between consecutive framework versions. We included substantive modifications to what the lab commits to do, additions of new commitments not present in prior versions,

removals of previously existing commitments, and meaningful alterations to the scope, strength, or enforceability of existing commitments².

When versions differed substantially in organization, as with Anthropic's transition from version 1.0 to 2.0, we tracked commitment trajectories rather than conducting section-by-section comparison. For each commitment in the source version, we identified its equivalent or absence in the target version. When classification was uncertain, we defaulted to the more conservative coding and documented the ambiguity in our notes.

Coding Scheme

Each identified change was coded across multiple dimensions. For direction, we classified changes as Strengthening when they increase the strength, scope, measurability, or enforceability of safety commitments; Weakening when they decrease these properties; or Neutral when the change is substantive but does not clearly shift commitment strength in either direction. Direction reflects the change's effect on external accountability, not the lab's stated intent, a change justified as "enabling flexibility" may still be coded as Weakening if it reduces verifiable commitments.

We developed an inductive taxonomy of mechanisms describing how commitments change, refined iteratively during coding. Weakening mechanisms include threshold inflation (raising the bar for triggering requirements), quantitative-to-qualitative shifts (replacing specific metrics with judgment-based criteria), discretion expansion (increasing internal decision-making latitude), accountability removal (reducing external oversight), scope narrowing (reducing coverage of risks or actors), timeline extension (permitting more time before action), definitional vagueness (making criteria less precise), and structural reset (reorganizing documents in ways that obscure comparison). Strengthening mechanisms include threshold lowering, commitment addition, specificity increase, scope expansion, and oversight addition. Complete definitions and examples appear in [Appendix A](#).

Each change was assigned to one of ten operational categories reflecting the type of commitment affected: thresholds, evaluations, deployment safeguards, security safeguards, internal oversight, external oversight, halt conditions, transparency, scope, and governance meta-processes. Full definitions appear in [Appendix B](#).

Two binary flags capture cross-cutting impact. The accountability flag indicates whether external parties' ability to verify or enforce compliance is affected. The measurability flag indicates whether commitments become more or less quantifiable. We also recorded whether each change was mentioned in the lab's official changelog, enabling analysis of transparency patterns.

Coding Process

For each version pair, we conducted systematic comparison using official changelogs as a starting point where available, followed by full-text comparison for

² We excluded formatting or stylistic changes, reorganization without substantive modification, clarifications that do not alter operative meaning, and definitions added solely to support another tracked change.

sections with substantive content. We employed AI-assisted initial identification of changes, followed by manual verification and classification. Changes were extracted with supporting quotes from both source and target versions, including page references. Ambiguous cases were flagged and resolved through explicit reasoning documented in notes.

Analytical Approach

Our analysis combines descriptive statistics with pattern identification. We calculated change counts by direction, lab, mechanism, and operational category, as well as changelog disclosure rates by direction. We identified cross-lab patterns such as universal changes to evaluation frequency, temporal patterns distinguishing major rewrites from incremental updates, and asymmetries in how weakenings versus strengthenings are disclosed.

We cannot verify whether changelogs capture all changes labs considered significant; absence from a changelog may reflect oversight rather than intent to obscure. We analyzed framework text, not implementation; whether labs follow their stated commitments is a separate research question. Finally, changes may reflect genuine learning, capability advances, regulatory pressure, or competitive dynamics; we do not attribute causation.

Reproducibility

The full dataset is publicly available, including all 62 coded changes with metadata, source quotes and page references, coder notes and reasoning for ambiguous cases, and links to all framework versions analyzed. The taxonomy and coding scheme are designed for application to future framework updates, enabling ongoing monitoring as frameworks evolve.

3. Results

Overall Distribution

Our analysis identified 62 discrete commitment changes across the three frameworks: 34 classified as Weakening (62%) and 21 as Strengthening (38%)³. This aggregate picture, however, obscures substantial variation across labs.

Divergent Lab Trajectories

The three labs exhibited markedly different patterns. OpenAI showed the most pronounced weakening, with 19 Weakening changes and zero Strengthening. Anthropic exhibited rough balance with 11 Weakening and 10 Strengthening changes. DeepMind showed net strengthening with only 4 Weakening changes against 11 Strengthening. These results appear in Figure 1.

³ Excluding neutral changes.

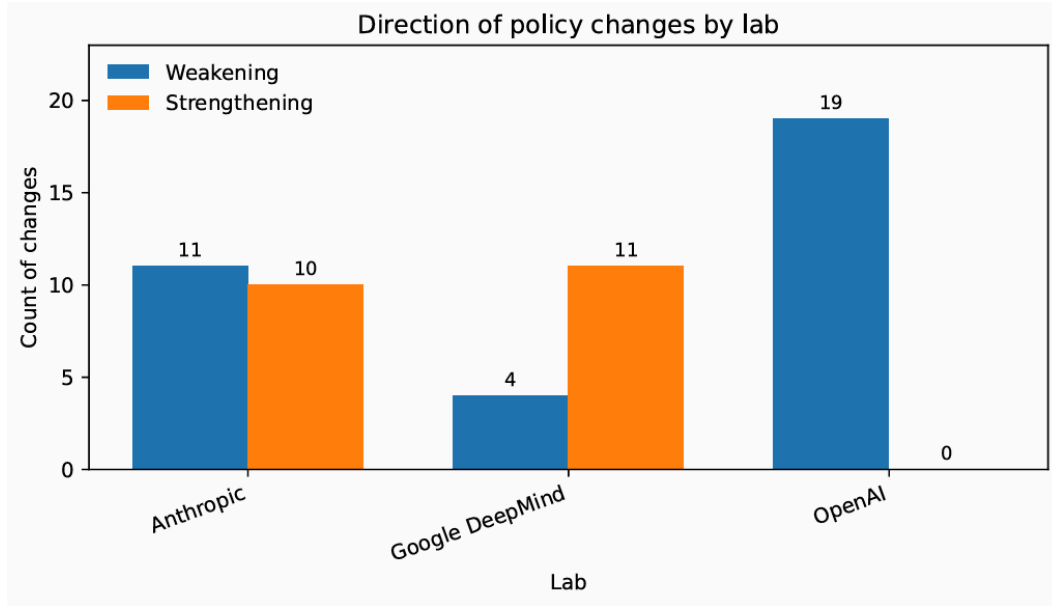


Figure 1: Changes by Direction per Lab

This divergence challenges simple narratives of uniform industry erosion. DeepMind demonstrates that framework updates can meaningfully strengthen commitments, while OpenAI's trajectory raises concerns about systematic weakening without compensatory improvements.

Mechanisms of Change

The most frequent weakening mechanisms were threshold inflation (8 instances), accountability removal (7 instances), quantitative-to-qualitative shifts (6 instances), and discretion expansion (5 instances). Among strengthening mechanisms, commitment addition (11 instances) and specificity increase (7 instances) dominated. The full distribution appears in Figure 2.

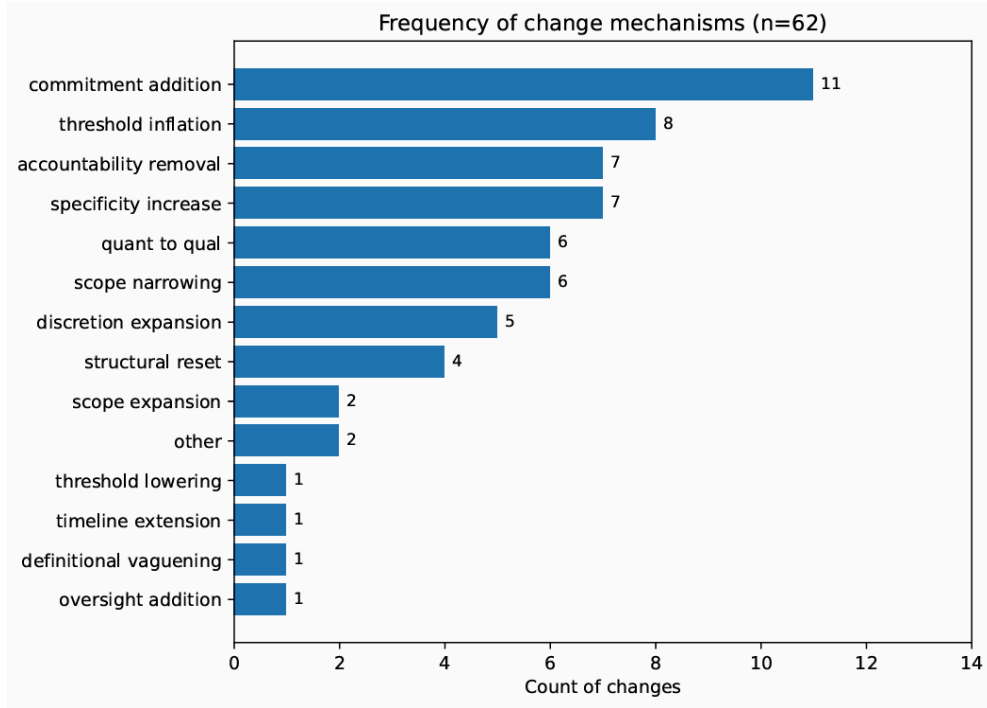


Figure 2: Distribution of Mechanisms

The prevalence of quantitative-to-qualitative shifts is particularly notable. Multiple labs replaced specific, measurable criteria with language requiring internal judgment: Anthropic's defined "6x safety buffer" became assessments of whether reaching thresholds is "unlikely," while DeepMind's "every 6x compute and every 3 months" became simply "regularly."

Changelog Transparency

We observed asymmetry in disclosure. 11 of 34 Weakening changes (32%) did not appear in official changelogs, while 13 of 21 Strengthening changes (62%) were not disclosed. Thus, strengthening changes were nearly twice as likely to be omitted as weakening changes, a pattern shown in Figure 3.

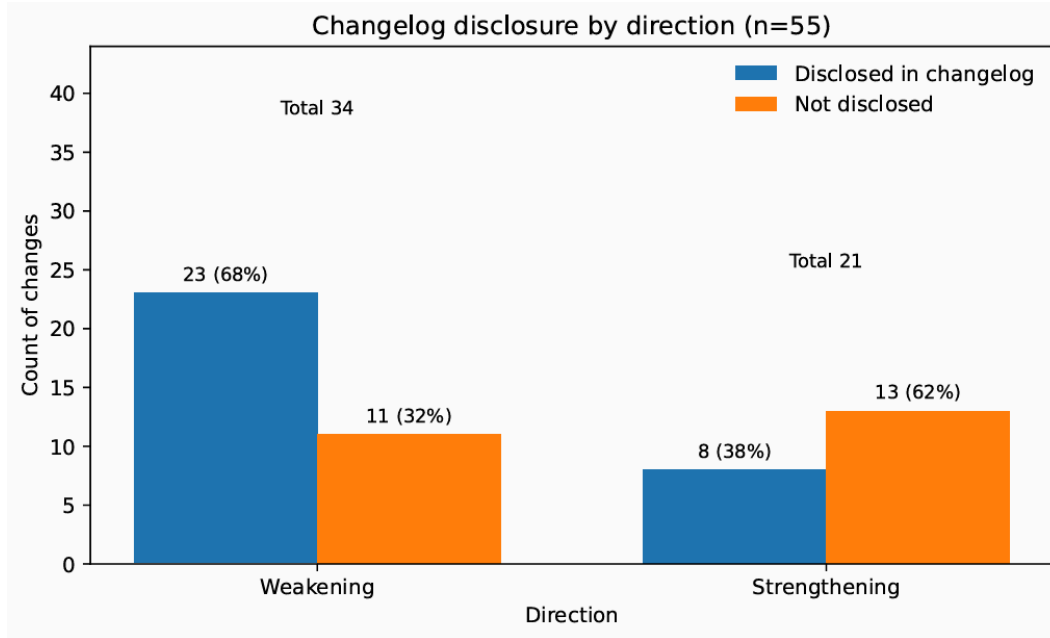


Figure 3: Changelog Disclosure Rates by Direction

4. Discussion and Conclusion

Limitations

Several limitations constrain interpretation. The sample of total 62 changes across three labs limits statistical inference; findings are indicative rather than definitive. Coding direction and mechanism involves judgment, and reasonable analysts may disagree on borderline cases. We mitigated this through explicit definitions, documented reasoning, and conservative defaults for ambiguous cases. Some early framework versions may be unavailable or were not formally versioned, so our analysis covers documented versions only.

Cross-Lab Patterns

Despite divergent trajectories, all three labs weakened evaluation frequency commitments. Anthropic extended intervals from three to six months. DeepMind replaced specific triggers with "regularly." None prominently disclosed these changes. The universal weakening of evaluation frequency is significant because regular evaluation is foundational to detecting capability advances before they cross dangerous thresholds.

Labs diverged on persuasion risks: OpenAI removed it from tracked categories arguing it requires "societal-level responses," while DeepMind added "Harmful Manipulation" as a new risk domain with defined capability levels.

Temporal Patterns

Comparing major and minor updates revealed that commitments added in one version may be walked back in subsequent versions. The governance bodies explicitly named in DeepMind's version 2.0 were replaced with generic "appropriate

corporate governance bodies" in version 3.0. Misalignment mitigations specified as required in version 2.0 became merely "illustrative" in version 3.0. Such reversals would be difficult to detect without systematic tracking.

Impact on Accountability

Among weakening changes (n=34), 23 (68%) affected external accountability and 19 (56%) affected measurability. Strengthening changes show similar rates (accountability: 13/21, 62%; measurability: 11/21, 52%). Overall, both strengthening and weakening edits frequently touch oversight-relevant properties, rather than being confined to internal operational details.

Policy Implications

Our results suggest that certain commitment types, especially evaluation frequency, external oversight, and quantitative thresholds, are poor candidates for purely voluntary maintenance, because even when labs update frameworks in good faith, these properties are repeatedly softened in ways that reduce verifiability and external accountability.

If policymakers want these safeguards to persist under competitive and organizational pressure, they should likely require minimum regulatory baselines (e.g., mandatory evaluation cadence, clearly specified trigger thresholds, and independent review/audit requirements for high-risk determinations).

Finally, the transparency gap is addressable: regulators could require mandatory disclosure of framework changes (a standardized changelog plus a machine-readable diff and rationale for any change affecting oversight-relevant properties), making quiet weakening harder and comparisons tractable.

Conclusion

This report provides a systematic taxonomy and version-to-version empirical analysis of how frontier AI safety frameworks evolve across Anthropic, OpenAI, and Google DeepMind. We coded 62 discrete commitment changes by direction, mechanism, affected commitment type, and whether the change was disclosed in official changelogs.

The results show divergence rather than uniform erosion: OpenAI's trajectory is dominated mostly by weakenings, Anthropic is roughly balanced, and DeepMind net strengthens. Transparency is also uneven, changelogs routinely miss substantive edits and one vulnerability is consistent across labs: evaluation-frequency commitments were weakened in all three frameworks without prominent disclosure, despite being foundational to detecting capability shifts.

Our dataset and taxonomy enable ongoing monitoring as frameworks update, turning "trust us" documents into trackable objects. If labs repeatedly soften evaluation frequency, external oversight, or quantitative thresholds, that's evidence these commitments may be difficult to sustain through voluntary policy alone. That strengthens the case for baseline requirements and clear disclosure rules for material changes.

5. References

METR. (2025a, December 9). Common elements of frontier AI safety policies (December 2025 update).
<https://metr.org/blog/2025-12-09-common-elements-of-frontier-ai-safety-policies>;

6. Appendix

Appendix A: Mechanism Taxonomy

Weakening Mechanisms

- **Threshold inflation** refers to changes that make safety requirement triggers harder to reach. For example, language shifting from "significant uplift" to "dramatic uplift" raises the bar for when safeguards apply.
- **Quantitative-to-qualitative** shift occurs when specific, measurable criteria are replaced with judgment-based assessments. An example is replacing a defined "6x safety buffer" with language stating the lab will assess whether reaching a threshold is "unlikely."
- **Discretion expansion** increases internal decision-making latitude by replacing fixed rules with processes requiring leadership determination. Automatic triggers that previously mandated specific responses become investigations that may or may not lead to action.
- **Accountability** removal reduces or eliminates external oversight mechanisms. Examples include removing requirements for third-party audits or independent review of safety assessments.
- **Scope narrowing** reduces the coverage of risks, actors, or scenarios addressed by the framework. This may involve excluding certain threat categories or carving out specific actor types from protection requirements.
- **Timeline extension** permits more time before required actions must be taken, such as extending evaluation intervals from three months to six months.
- **Definitional vagueness** makes criteria less precise, for example replacing specific capability lists with references to unspecified "capabilities of concern."
- **Structural reset** involves document reorganization that makes comparison difficult, potentially obscuring whether previous commitments have been maintained.

Strengthening Mechanisms

- **Threshold lowering** makes safety protocols trigger earlier, such as reducing an acceleration threshold from 3x to 2x.
- **Commitment addition** introduces entirely new obligations not present in prior versions.
- **Specificity** increase makes requirements more precise or measurable, such as replacing "periodic" review with "at least annual" review.

- **Scope expansion** increases the coverage of risks, actors, or scenarios, such as expanding from biosecurity to full CBRN coverage.
- **Oversight addition** introduces new external accountability mechanisms such as third-party procedural audits.

Appendix B: Operational Categories

- **Thresholds** encompasses capability threshold definitions including ASL levels, CCLs, and risk tiers that determine when enhanced safeguards apply.
- **Evaluations** covers evaluation frequency, scope, methodology, and elicitation techniques used to assess model capabilities.
- **Deployment safeguards** includes safeguards applied at deployment time such as red-teaming requirements, usage restrictions, and mitigation measures.
- **Security safeguards** addresses security standards, weight protection measures, and insider threat controls.
- **Internal oversight** covers internal governance mechanisms including board approval requirements, safety team authority, and decision-making processes.
- **External oversight** encompasses third-party audits, independent review requirements, and government disclosure commitments.
- **Halt conditions** defines the conditions that trigger pausing or halting training or deployment activities.
- **Transparency** covers public disclosure requirements including model cards, incident reporting, and evaluation result publication.
- **Scope** addresses what models, systems, or risks fall within the framework's coverage.
- **Governance meta** describes how the framework itself is updated, amended, or governed over time.