Standards Body · Research note, public edition · Released July 17, 2026
Canonical record: https://standardsbody.ai/library/research-note/frontier-framework-crosswalk/

FRONTIER_FRAMEWORK_CROSSWALK.md

A Crosswalk of Frontier AI Safety Frameworks: Evaluation Triggers and Commitments

Project: Standards Body **Primary domain:** standardsbody.ai **Core line:** Foundations for Frontier AI **Document type:** Comparative research note **Version:** 1.0 **Status:** Research note. Noncanonical. Authority effect: None. **Research basis reviewed through:** July 18, 2026 **Review state:** Not externally reviewed. Framework publishers are explicitly invited to correct any characterization of their documents.

Document Purpose

This research note compares the three published frontier AI safety frameworks that currently anchor voluntary risk governance at the leading AI developers: Anthropic's Responsible Scaling Policy, OpenAI's Preparedness Framework, and Google DeepMind's Frontier Safety Framework. It documents what each framework covers, what triggers evaluation under each, what crossing a threshold commits the developer to, and where the frameworks' structures and vocabularies diverge.

The comparison exists because these three documents are, at present, the closest thing the field has to operating evaluation-triggered governance, and because they are routinely discussed as if they were parallel instruments when they are structurally different in ways that matter. A policy analyst who reads that a model "did not cross a threshold" needs to know that the word threshold binds three different things in these three frameworks.

Method and Limitations

This crosswalk is built exclusively from the publishers' own public documents and announcements, listed in the Sources section with access dates. It reflects those documents as of July 18, 2026. All three frameworks describe themselves as living documents and change frequently; any characterization here may be outdated by later revisions, and this note will be corrected or superseded accordingly.

In Standards Body terminology, this note is a structured comparison of published commitments. It is not an audit, which would require defined criteria and independence; not a conformity assessment; not an evaluation of any model; and not a claim about whether any developer follows its framework in practice. Compliance is outside its scope entirely. Descriptions of what a framework requires are descriptions of the document, not attestations about behavior.

Each characterization was drafted from the source text and is stated in this note's own words. Where a framework's self-description and third-party summaries conflicted, the publisher's document was used. Any framework publisher who believes a characterization here is inaccurate is invited to say so; corrections are recorded and released versions are preserved. The correction route is at the end of this note.

The Three Frameworks at a Glance

	Anthropic Responsible Scaling Policy	**OpenAI Preparedness Framework**	**Google DeepMind Frontier Safety Framework**
Current version	3.0, effective February 24, 2026, a comprehensive rewrite	2, April 15, 2025	3, September 22, 2025, with Tracked Capability Levels added as of April 17, 2026
First published	September 2023	December 2023 (beta)	May 2024
Central threshold concept	Capability or usage thresholds mapped to mitigations, within industry-wide recommendations and company plans	Two capability thresholds per Tracked Category: High and Critical	Critical Capability Levels (CCLs) per risk domain, with earlier Tracked Capability Levels (TCLs) in some domains
Central evaluation output	Risk Reports published every 3 to 6 months, covering capabilities, threat models, mitigations, and an overall risk determination	Capability evaluations against threshold definitions, governed by a Safety Advisory Group	Early-warning evaluations against alert thresholds with safety buffers, plus holistic risk assessments; FSF reports published for frontier releases
Distinctive structural feature	Separates unilateral company commitments from industry-wide recommendations, with competitor-contingent commitments	Explicit two-tier deployment and development gating per category	Alert thresholds and safety buffers designed to trigger review before a critical level is reached

---

Domain Coverage Compared

The frameworks agree on a core and diverge at the edges. Reading the domain lists side by side shows where each publisher believes evaluation-triggered governance currently belongs.

Risk domain	**Anthropic RSP v3.0**	**OpenAI PF v2**	**DeepMind FSF v3**
	Covered, split into two thresholds: uplift to basic-		Covered within the CBRN domain

Risk domain	Anthropic RSP v3.0	OpenAI PF v2	DeepMind FSF v3
Chemical and biological weapons uplift	background individuals or groups (non-novel), and uplift to expert-backed teams toward novel weapons	Covered as one Tracked Category: Biological and Chemical	
Nuclear and radiological	Within the chemical and biological rows the RSP addresses chemical and biological weapons specifically	A Research Category, not yet a Tracked Category	Included in the CBRN domain
Cyber operations	Not one of the four named thresholds in v3.0's recommendations table	Covered as a Tracked Category: Cybersecurity	Covered as a domain with its own CCL and, in current reporting, an earlier alert threshold
AI research automation and self-improvement	Covered: automated R&D in key domains, operationalized as the ability to compress two years of 2018 to 2024 AI progress into one year	Covered as a Tracked Category: AI Self-Improvement	Covered as the machine learning R&D domain, assessed by automation levels
Internal sabotage, deception, misalignment	Covered: high-stakes sabotage opportunities for AI systems with extensive access and autonomous operation	Related areas appear as Research Categories: Long-range Autonomy, Sandbagging, Autonomous Replication and Adaptation, Undermining Safeguards	Addressed as misalignment risk, described by the publisher as exploratory, with instrumental reasoning levels
Manipulation and persuasion	Not a named threshold; earlier versions noted persuasive capability as insufficiently understood to include	Explicitly handled outside the framework, through other company mechanisms	Covered since v3: a harmful manipulation CCL for systematic belief and behavior change in high-stakes contexts

Three observations follow directly. First, only DeepMind currently treats manipulation as an evaluation-triggering domain; OpenAI expressly routes it elsewhere and Anthropic does not name it. Second, cyber capability is a first-class tracked domain for OpenAI and DeepMind but is not one of the four named thresholds in Anthropic's v3.0 recommendations table. Third, all three converge on AI research automation as a threshold domain, and all three now operationalize it with distinct measures: a progress-compression definition (Anthropic), a Tracked Category (OpenAI), and automation levels with benchmark-based leading indicators (DeepMind).

What Triggers Evaluation

Question	Anthropic RSP v3.0	OpenAI PF v2	DeepMind FSF v3
When are evaluations or assessments due?	Risk Reports every 3 to 6 months covering all publicly deployed models; additional analysis published when a significantly more capable model is deployed,	Evaluations are tied to the framework's threshold assessments for covered systems, with threat models and thresholds reviewed and	Early-warning evaluations run at a regular cadence and after significant capability jumps

Question	Anthropic RSP v3.0	OpenAI PF v2	DeepMind FSF v3
	and within 30 days of an in-scope internal deployment	approved by the Safety Advisory Group	
Are internal, non-released models in scope?	Yes, when determined to pose significant risks beyond public models, explicitly including models internally deployed for large-scale fully autonomous research	The framework applies to covered systems under its capability criteria	Yes; assessments address models before external deployment
Is there a pre-threshold warning layer?	Earlier versions used a checkpoint concept (for example, a software engineering task horizon); v3.0 notes earlier, easier-to-measure thresholds may be sensible	No formal pre-threshold tier below High	Yes, twice over: alert thresholds with safety buffers below each CCL, and, since April 2026, Tracked Capability Levels for earlier, less extreme risk

The historical direction of travel matters as much as the current state. Anthropic's own changelog records a movement from prespecified quantitative evaluations (v1.0), to affirmative capability cases with a cadence of 4x effective compute or six months of accumulated post-training enhancements (v2.0 to v2.2), to the v3.0 structure organized around periodic Risk Reports and argument-based determinations rather than AI Safety Levels defined in advance. DeepMind has moved in the opposite direction on granularity, adding a manipulation CCL and then a below-critical tracked tier. OpenAI simplified from four levels to two. The three frameworks are not converging on one evaluation-trigger architecture.

What Crossing a Threshold Commits the Developer To

	Anthropic RSP v3.0	OpenAI PF v2	DeepMind FSF v3
At the primary threshold	The company states planned mitigations per threshold (for example, maintaining or improving ASL-3-class protections for chemical and biological uplift) and publishes its determinations and reasoning in Risk Reports	High capability: the system must have safeguards that sufficiently minimize the associated risk before deployment, plus security controls	Reaching a CCL triggers the framework's mitigation approach: security mitigations addressing model weight protection and deployment mitigations restricting access to critical capability, with a risk-acceptability determination
At the higher tier	For automated R&D, the highly capable determination triggers, among other things, mandatory external review of significantly redacted Risk Reports	Critical capability: safeguards are required during development itself, not only before deployment	CCLs are the critical tier; TCLs sit below them and trigger monitoring and evaluation rather than critical-level mitigations
Development pause posture	The RSP commits to specific scenarios: if clearly in the lead on a highly capable model, delay development as needed until a strong safety argument exists;	The two-tier structure is the framework's gating mechanism: deployment gated at High,	The framework's posture is that assessed capability must be found acceptable before external deployment proceeds

	Anthropic RSP v3.0	OpenAI PF v2	DeepMind FSF v3
	competitor-contingent commitments otherwise	development safeguards at Critical	

A reader should not treat these as three strengths of the same commitment. They are different commitment types. OpenAI's is a conditional gate stated per category. DeepMind's is an evaluation-and-mitigation pipeline with the determination step made explicit. Anthropic's v3.0 is, by its own account, a change in kind: it separates what the company will do unilaterally from what it recommends industry-wide, and makes some commitments contingent on competitor behavior, on the stated reasoning that unilateral pauses by responsible developers could make the ecosystem less safe. Whether that reasoning is right is a substantive question this note does not adjudicate; that the commitment structure changed is a documented fact a comparison must surface.

External Review and Transparency Compared

All three frameworks publish framework documents, model-level safety reporting, and change logs. They differ on independent review. Anthropic's v3.0 commits to working toward comprehensive public external review of Risk Reports, with defined reviewer independence criteria (no financial interest, no close personal relationships in the review chain) and a mandatory external review when a report covers highly capable models and is significantly redacted; it also commits to an approximately annual third-party review of procedural compliance. OpenAI's framework centers internal governance through its Safety Advisory Group. DeepMind's framework reports describe structured internal risk assessment with defined decision owners, and its frontier releases have involved pre-deployment evaluations by government AI safety and security institutes.

From the standpoint of Standards Body's terminology, none of these arrangements yet constitutes audit or certification: there is no common external criteria set, no accredited assessing body, and no cross-developer comparability of conclusions. The strongest present mechanisms are developer-selected external review and government institute testing, both meaningful, neither equivalent to independent conformity assessment. That is an observation about the state of the ecosystem's assurance infrastructure, not a criticism of any single developer, since no such infrastructure currently exists for anyone to use.

Terminology Divergence

The same words do different work in the three frameworks, and some different words do the same work. The mappings below use the definitions in Standards Body Terminology.

Threshold. Anthropic's capability or usage thresholds, OpenAI's High and Critical capability thresholds, and DeepMind's Critical Capability Levels are all prespecified capability levels that trigger obligations, and in that sense are comparable. But Anthropic's v3.0 thresholds live inside recommendations and plans rather than an if-then gating table; OpenAI's are two tiers within each of three categories; DeepMind's are per-domain levels with a formal sub-critical tier. A sentence

like "the model did not cross any thresholds" is therefore underdetermined without naming the framework.

Evaluation. All three use evaluation broadly in line with its proper meaning: a structured process for producing and interpreting evidence, not merely test execution. DeepMind's early-warning evaluations and alert thresholds make the interpretive layer explicit; Anthropic's Risk Reports fold evaluation results into an overall risk argument; OpenAI ties evaluations to threshold determinations under governance review.

Safety level. Anthropic's ASL now denotes groups of present risk mitigations, not model categories, and v3.0 explicitly steps away from defining future requirements as levels. Treating ASL as a model rating, as public discussion often does, misreads the current document.

Assurance. None of the three frameworks claims absolute assurance, and each contains language acknowledging that risk cannot be eliminated. Public summaries that render a framework determination as "the model is safe" convert a bounded, dated, self-assessed conclusion into a universal claim that no framework makes.

Observations

Four things emerge from reading the documents side by side rather than through summaries.

First, structural comparability is eroding, not improving. Between 2024 and 2026 the three frameworks moved further apart in architecture: one toward periodic argument-based risk reporting with ecosystem-contingent commitments, one toward a simplified two-tier gate, one toward finer-grained tiers and added domains. Anyone building evaluation infrastructure, benchmarks, registries, or policy that assumes the frameworks are parallel instruments is building on an assumption the documents no longer support.

Second, the evaluation trigger is becoming a judgment, not a measurement, in at least one framework. The movement from prespecified evaluations to affirmative cases and risk arguments trades verifiability for adaptability. The publishers give reasons for this trade; the consequence for external observers is that framework adherence becomes progressively harder to assess from outside, which raises rather than lowers the value of independent review infrastructure.

Third, domain selection is a live disagreement. Manipulation, cyber, and nuclear or radiological risk are each treated as evaluation-triggering by some frameworks and not others. These are substantive judgments about threat models, made visible only by comparison.

Fourth, every framework now contains an explicit acknowledgment that single-developer commitments cannot secure the ecosystem, expressed as competitor-contingent commitments, calls for harmonized governance, or descriptions of frontier safety as a global public good. The frameworks themselves, in other words, argue for the existence of the third-party evaluation, standards, and assurance infrastructure whose foundations this project studies. That convergence, from three competing developers, is the most significant single finding of this comparison.

What This Crosswalk Does Not Do

It does not assess whether any developer complies with its framework. It does not rank the frameworks. It does not evaluate any model. It does not cover developers beyond these three, and notable frontier developers publish no comparable framework, which is itself relevant context. It does not capture unpublished internal practice, which may exceed or fall short of the documents. It reflects the sources as of July 18, 2026 and will age.

Sources

All sources accessed July 18, 2026.

1. Anthropic, Responsible Scaling Policy, Version 3.0, effective February 24, 2026. https://www.anthropic.com/responsible-scaling-policy/rsp-v3-0
2. Anthropic, Responsible Scaling Policy overview and update page. https://www.anthropic.com/responsible-scaling-policy
3. Anthropic, Responsible Scaling Policy, Version 2.2, effective May 14, 2025 (for historical thresholds and cadence). https://www.anthropic.com/rsp-updates
4. OpenAI, Preparedness Framework, Version 2, April 15, 2025. https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf
5. OpenAI, Our updated Preparedness Framework, April 2025. https://openai.com/index/updating-our-preparedness-framework/
6. Google DeepMind, Strengthening our Frontier Safety Framework, September 22, 2025, with the April 17, 2026 Tracked Capability Levels update. https://deepmind.google/blog/strengthening-our-frontier-safety-framework/
7. Google DeepMind, Frontier Safety Framework, Version 2.0, February 4, 2025 (for domain definitions and mitigation structure). https://deepmind.google/frontier-safety-framework/
8. Google DeepMind, Gemini 3 Pro Frontier Safety Framework Report, November 2025 (for evaluation practice, alert thresholds, and automation-level assessment). https://storage.googleapis.com/deepmind-media/gemini/gemini_3_pro_fsf_report.pdf

---

Corrections and Publisher Responses

Corrections: contact@standardsbody.ai. Corrections, supersession, and withdrawal are recorded; released versions are preserved without silent edits. Anthropic, OpenAI, and Google DeepMind are explicitly invited to correct any characterization of their documents in this note; publisher corrections will be prioritized, recorded, and attributed if desired.