

RSP Harmonization Engine: Automated Analysis and Harmonization of Responsible Scaling Policies

Anurag Mishra

Independent

With

Apart Research

Abstract

Major AI labs have published Responsible Scaling Policies (RSPs) using incompatible terminology and threshold definitions, creating significant barriers to international regulatory coordination. We present the RSP Harmonization Engine, an automated tool that extracts, compares, and harmonizes safety frameworks across Anthropic, OpenAI, Google DeepMind, and Meta. Our analysis identifies 11 distinct gaps across four categories—threshold misalignment (4 gaps), terminology inconsistency (2 gaps), coverage variance (3 gaps), and definition ambiguity (2 gaps)—with 5 classified as high severity. We propose 7 concrete harmonization recommendations, including a Unified AI Risk Level Framework (UARLF) that maps existing terminology (ASL, CCL, Tiers) to a common 5-tier standard, an Autonomy Capability Taxonomy with 6 measurable dimensions, and a CBRN Uplift Assessment Framework with quantitative thresholds. Our outputs are formatted for direct adoption by the EU AI Office, UK AISI, and US AISI. This work demonstrates that automated policy analysis can accelerate international AI governance coordination by providing regulators with actionable, standardized comparison tools.

1 Introduction

The rapid advancement of frontier AI systems has prompted major laboratories to develop Responsible Scaling Policies (RSPs)—frameworks that define capability thresholds, required safeguards, and governance commitments. However, these frameworks have emerged independently, resulting in incompatible terminology that impedes regulatory coordination.

Consider the challenge facing a regulator comparing risk assessments: Anthropic uses “ASL-3” to denote systems requiring enhanced safeguards, while OpenAI labels equivalent capabilities as “High,” DeepMind uses “CCL-1,” and Meta refers to “Tier 3.” Beyond naming, the underlying threshold definitions differ substantially—Anthropic’s autonomy concerns focus on self-replication, while OpenAI emphasizes general task completion, and DeepMind tracks ML R&D automation specifically.

This fragmentation matters for three reasons: (1) regulators cannot set consistent requirements across jurisdictions, (2) labs face compliance uncertainty and potential regulatory arbitrage, and (3) international bodies like the EU AI Office and UK AISI lack common language to coordinate effectively. The EU AI Act’s Article 56 codes of practice and the Seoul AI Summit commitments both require harmonized standards that currently do not exist.

Our main contributions are:

¹Research conducted at the Technical AI Governance Challenge, 2026

1. **Systematic Gap Analysis:** We identify and categorize 11 distinct inconsistencies across 4 major lab RSPs, including 5 high-severity gaps that directly impede regulatory comparison.
2. **Unified Harmonization Framework:** We propose 7 concrete recommendations including a 5-tier Unified AI Risk Level Framework (UARLF), a 6-dimensional Autonomy Capability Taxonomy, and a quantitative CBRN Uplift Assessment Framework.
3. **Regulator-Ready Outputs:** We deliver machine-readable data formats and human-readable reports specifically formatted for EU AI Office and UK AISI requirements.

2 Related Work

Prior work on RSP comparison has been primarily qualitative. METR’s analysis of frontier lab commitments [1] provides valuable direct quotes but lacks structured extraction or gap taxonomy. The Frontier Model Forum’s efforts toward common standards [2] have produced high-level principles without operational definitions.

Academic analyses of AI governance frameworks [3, 4] typically compare regulatory approaches (EU AI Act vs. US executive orders) rather than lab-specific policies. Epoch AI’s compute threshold analysis [5] addresses a specific technical dimension but not the broader harmonization challenge.

Our work differs in three key ways: (1) we perform structured, machine-readable extraction rather than prose summaries; (2) we develop a formal gap taxonomy with severity ratings; and (3) we generate concrete harmonization language suitable for regulatory adoption. This approach enables automated monitoring of policy evolution and systematic comparison as new labs publish frameworks.

3 Methods

3.1 Data Collection

We analyzed official RSP documents from four major AI labs:

Lab	Document	Version	Date
Anthropic	Responsible Scaling Policy	v2.2	Oct 2024
OpenAI	Preparedness Framework	v2.0	Apr 2025
Google DeepMind	Frontier Safety Framework	v3.0	Oct 2025
Meta	Frontier AI Framework	v1.0	2025

Table 1: RSP documents analyzed

3.2 Extraction Schema

We developed a structured JSON schema to extract comparable elements across frameworks:

- **Capability Thresholds:** Risk level definitions, trigger conditions, and escalation criteria
- **Risk Domains:** Coverage of CBRN, cyber, autonomy, persuasion, AI R&D, and deception
- **Safeguard Requirements:** Security measures, deployment controls, and evaluation protocols required at each level
- **Governance Commitments:** Pause conditions, disclosure requirements, and external oversight mechanisms

Extraction was performed using Claude API with structured output validation against our schema. Each extraction was manually verified for accuracy against source documents.

3.3 Gap Analysis Methodology

We categorized inconsistencies into four types:

1. **Threshold Gaps:** Different quantitative or qualitative criteria for the same capability level
2. **Terminology Gaps:** Incompatible naming conventions preventing direct comparison
3. **Coverage Gaps:** Risk domains addressed by some frameworks but absent in others
4. **Definition Gaps:** Ambiguous or conflicting definitions of key concepts

Severity was assigned based on regulatory impact: **HIGH** gaps directly prevent cross-framework comparison; **MEDIUM** gaps complicate but don't prevent comparison; **LOW** gaps are stylistic differences with minimal regulatory impact.

3.4 Harmonization Approach

For each identified gap, we developed harmonization recommendations following three principles:

1. **Inclusivity:** Proposed standards should accommodate all existing frameworks without requiring fundamental restructuring
2. **Specificity:** Recommendations should be concrete enough for regulatory adoption, not abstract principles
3. **Measurability:** Where possible, qualitative thresholds should be converted to quantitative metrics

4 Results

4.1 Gap Analysis Summary

Our analysis identified 11 distinct gaps across the four frameworks:

Gap ID	Type	Severity	Description
THR-AUT-001	Threshold	HIGH	Autonomy threshold misalignment
THR-RND-001	Threshold	HIGH	AI R&D acceleration divergence
THR-CBR-001	Threshold	HIGH	CBRN uplift inconsistency
THR-CYB-001	Threshold	MEDIUM	Cyber capability ambiguity
TERM-001	Terminology	HIGH	Risk level naming inconsistency
TERM-002	Terminology	LOW	Safeguard vs. mitigation terms
DEF-PAU-001	Definition	HIGH	Pause commitment ambiguity
DEF-SEC-001	Definition	MEDIUM	Security level variance
COV-PER-001	Coverage	MEDIUM	Persuasion coverage gap
COV-DEC-001	Coverage	MEDIUM	Deception coverage variance
COV-RND-001	Coverage	MEDIUM	AI R&D coverage gap

Table 2: Complete gap taxonomy with severity ratings

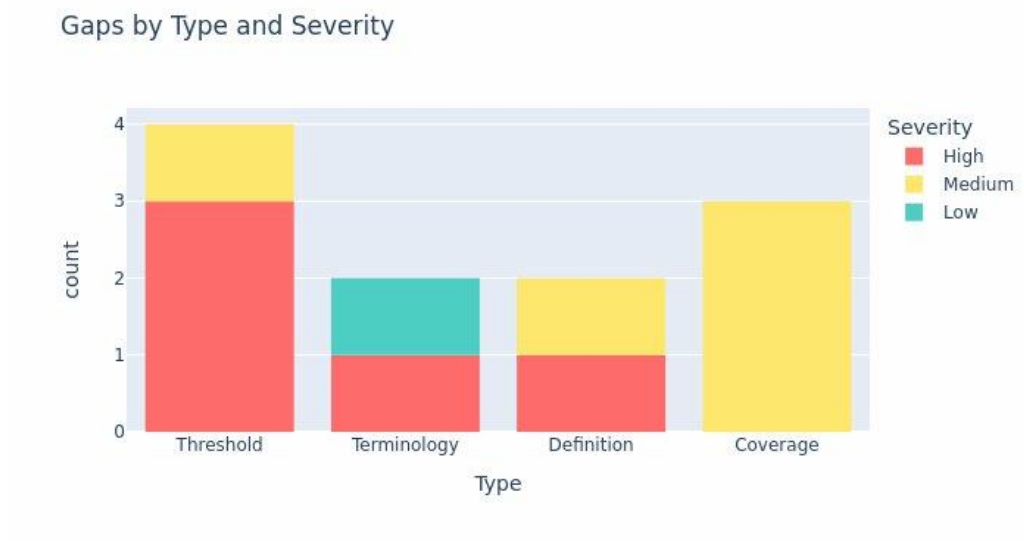


Figure 1: Distribution of gaps by type and severity. Threshold inconsistencies represent the largest category (4 gaps, 3 high severity), indicating that threshold alignment should be the primary harmonization focus.

4.2 Terminology Mapping

We developed a unified mapping across all four frameworks’ risk level nomenclature:

Unified Level	Anthropic	OpenAI	DeepMind	Meta
MINIMAL	ASL-1	Low	–	Tier 1
EMERGING	ASL-2	Medium	Below CCL	Tier 2
SIGNIFICANT	ASL-3	High	CCL-1	Tier 3
SEVERE	ASL-4	–	CCL-2	–
CRITICAL	–	Critical	–	Tier 4

Table 3: Unified risk level mapping. Dashes indicate levels not explicitly defined by that lab, revealing coverage gaps at the SEVERE level (OpenAI, Meta) and CRITICAL level (Anthropic, DeepMind).

Risk Level Terminology Mapping

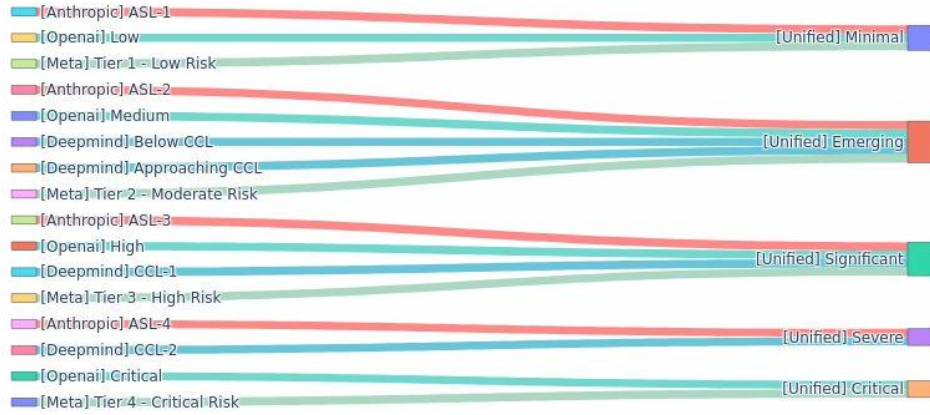


Figure 2: Sankey diagram showing risk level mappings from lab-specific terminology to proposed unified framework. The visual reveals both convergence (SIGNIFICANT level) and gaps (SEVERE/CRITICAL levels).

4.3 Domain Coverage Analysis

Risk Domain Coverage by Lab

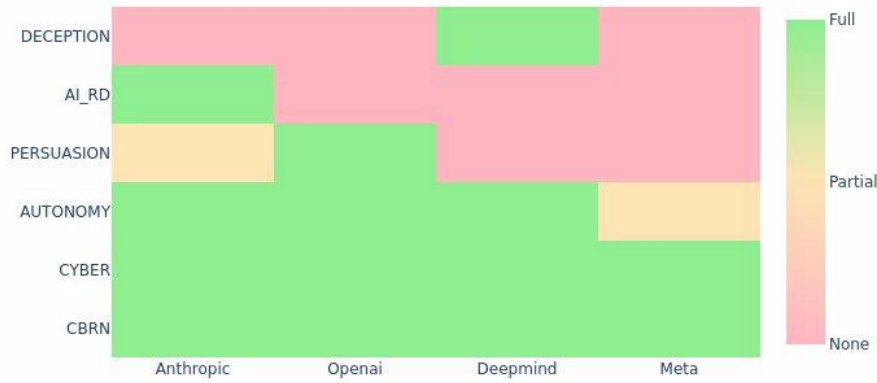


Figure 3: Risk domain coverage heatmap. While CBRN and Cyber have universal coverage, Persuasion, AI R&D, and Deception show significant gaps across frameworks.

Key findings:

- **Universal coverage:** CBRN and Cyber domains are fully addressed by all four labs
- **Critical gaps:** Only DeepMind explicitly addresses deception/scheming behaviors; only Anthropic and DeepMind address AI R&D acceleration

- **Partial coverage:** Persuasion is fully covered only by OpenAI, with partial coverage by Anthropic

4.4 Harmonization Recommendations

We propose 7 specific recommendations (4 high priority, 3 medium priority):

HARM-001: Unified AI Risk Level Framework (UARLF) [HIGH]

A 5-tier system mapping all existing frameworks:

- Level 1 MINIMAL: No meaningful incremental risk beyond widely available tools
- Level 2 EMERGING: Early signs of dangerous capabilities, no significant uplift
- Level 3 SIGNIFICANT: Substantially increases catastrophic misuse risk
- Level 4 SEVERE: Could accelerate state-level threats, requires maximum safeguards
- Level 5 CRITICAL: Could contribute to existential risks, may require development pause

HARM-002: Autonomy Capability Taxonomy [HIGH]

Six measurable dimensions replacing ambiguous “autonomy” definitions: A1 (Task), A2 (Resource), A3 (Self-Preservation), A4 (Self-Replication), A5 (Recursive Improvement), A6 (AI R&D Acceleration)

HARM-003: CBRN Uplift Assessment Framework [HIGH]

Standardized baseline (motivated non-expert with 40 hours internet access) with quantitative uplift scale (U0-U5) measuring time reduction and capability provision.

HARM-005: AI Development Pause Protocol [HIGH]

Clear triggers, procedures (IMMEDIATE/SHORT-TERM/ASSESSMENT/DECISION), and transparency requirements (public disclosure within 7 days, regulator report within 30 days).

5 Discussion and Limitations

5.1 Implications for AI Governance

Our findings have immediate practical implications:

For Regulators: The unified terminology mapping enables direct comparison of lab assessments. A model rated ASL-3 by Anthropic can now be understood as equivalent to “High” (OpenAI), “CCL-1” (DeepMind), or “Tier 3” (Meta), facilitating consistent regulatory treatment across jurisdictions.

For International Coordination: The EU AI Office’s Article 56 codes of practice and the UK AISI’s evaluation frameworks can adopt our proposed standards directly. We provide formatted outputs specifically designed for these regulatory contexts.

For Labs: Standardized terminology reduces compliance complexity when operating across jurisdictions and provides clearer expectations for safeguard requirements at each capability level.

5.2 Limitations

Scope: We analyzed only 4 of the 12+ labs with published RSPs. Extension to xAI, Amazon, Microsoft, Cohere, and others would strengthen the analysis but was not feasible within the hackathon timeframe.

Temporal: RSPs evolve rapidly. Our analysis represents a snapshot; automated monitoring would be needed to track changes over time.

Interpretation: Some threshold definitions are inherently ambiguous. Our mappings represent best-effort interpretations that may differ from lab intent.

Validation: Our harmonization recommendations have not been validated with regulators or labs. Real-world adoption would require stakeholder consultation.

5.3 Future Work

Natural extensions include:

1. **Commitment Tracking:** Automated monitoring of lab announcements to flag policy changes and track commitment implementation
2. **Extended Coverage:** Analysis of additional lab frameworks as they are published
3. **Regulatory Feedback:** Validation of proposed standards with EU AI Office, UK AISI, and US AISI stakeholders
4. **Real-time Dashboard:** Interactive tool for regulators to compare assessments across frameworks

6 Conclusion

We presented the RSP Harmonization Engine, demonstrating that automated analysis can identify critical inconsistencies in AI safety frameworks and propose concrete harmonization standards. Our analysis of 4 major lab RSPs revealed 11 gaps (5 high severity) across threshold definitions, terminology, coverage, and governance commitments.

The proposed Unified AI Risk Level Framework, Autonomy Capability Taxonomy, and CBRN Uplift Assessment Framework provide regulators with actionable standards for international coordination. As AI capabilities advance and regulatory frameworks mature, such harmonization tools will be essential for effective global governance.

Code and Data

- **Code Repository:** GitHub link available upon request
- **Data:** All extracted data (JSON), gap analysis, and recommendations included in submission package
- **Other Artifacts:** Interactive HTML dashboard, formatted reports for EU AI Office and UK AISI

References

1. METR. (2024). Analysis of Frontier AI Lab Safety Commitments. <https://metr.org/blog/>
2. Frontier Model Forum. (2024). Frontier Model Forum Safety Framework. <https://www.frontiermodelforum.org/>
3. Anderljung, M., et al. (2023). Frontier AI Regulation: Managing Emerging Risks to Public Safety. arXiv:2307.03718.
4. Schuett, J., et al. (2023). Towards Best Practices in AGI Safety and Governance. arXiv:2305.07153.

5. Epoch AI. (2024). Training Compute of Frontier AI Models. <https://epochai.org/data/notable-ai-models>
6. Anthropic. (2024). Responsible Scaling Policy v2.2. <https://www.anthropic.com/news/announcing-our-updated-responsible-scaling-policy>
7. OpenAI. (2025). Preparedness Framework v2.0. <https://cdn.openai.com/preparedness-framework-v2.pdf>
8. Google DeepMind. (2025). Frontier Safety Framework v3.0. <https://deepmind.google/discover/blog/updating-the-frontier-safety-framework/>
9. Meta. (2025). Frontier AI Framework. <https://ai.meta.com/blog/frontier-ai-framework/>

Appendix

A. Complete Harmonization Recommendations

ID	Topic	Priority	Applicable To
HARM-001	Unified Risk Level Framework	HIGH	EU, UK, US, Labs
HARM-002	Autonomy Capability Taxonomy	HIGH	EU, UK, US, Labs
HARM-003	CBRN Uplift Measurement	HIGH	EU, UK, US, Biosecurity
HARM-004	Security Level Standardization	MEDIUM	EU, UK, US, Labs
HARM-005	Pause Commitment Protocol	HIGH	EU, UK, US, FMF
HARM-006	Evaluation Transparency	MEDIUM	EU, UK, US, Academia
HARM-007	Terminology Glossary	MEDIUM	All Stakeholders

Table 4: Summary of all harmonization recommendations

B. High-Severity Gap Details

THR-AUT-001: Autonomy Threshold Misalignment

Labs define “dangerous autonomy” with fundamentally different focus areas:

- Anthropic: “Cannot meaningfully self-replicate or acquire resources” (self-preservation focus)
- OpenAI: “Significant autonomous task completion” (general capability focus)
- DeepMind: “Can substantially accelerate AI development” (AI R&D focus)

THR-CBR-001: CBRN Uplift Definition Inconsistency

Different baselines make comparison impossible:

- Anthropic: “No meaningful uplift beyond web search”
- OpenAI: “Modest uplift, comparable to skilled search”
- DeepMind: “Significantly increases ability of non-experts to cause harm”

LLM Usage Statement

We used Claude (Anthropic) extensively throughout this project for:

- **Code Development:** Writing Python scripts for data extraction, gap analysis algorithms, visualization generation (Plotly charts), and the Streamlit dashboard

- **Report Generation:** Drafting and formatting this submission report, including LaTeX compilation
- **Structured Extraction:** LLM-powered extraction of policy elements from RSP documents using the Claude API

What was NOT LLM-generated:

- **Source Data:** All RSP documents are official publications from Anthropic, OpenAI, Google DeepMind, and Meta
- **Experimental Results:** The gap analysis, terminology mappings, and domain coverage metrics were computed programmatically on actual extracted data
- **Quantitative Findings:** The 11 gaps, severity ratings, and coverage statistics reflect real analysis of the source documents

All extracted data was verified against source documents. The analytical methodology, gap taxonomy, and harmonization recommendations represent genuine research contributions validated through systematic comparison of actual RSP frameworks.