

AI Safety Threshold Tracker

An Interactive Dashboard for Monitoring AI Capabilities Against Governance Frameworks

Revathi Prasad

Independent Researcher

With **APART Research**

Abstract

As AI capabilities rapidly advance, policymakers and researchers lack unified tools connecting capability benchmarks to governance-defined thresholds. We present the AI Safety Threshold Tracker, an interactive dashboard that aggregates verified capability data from two categories of authoritative sources: (1) third-party benchmarks (METR, SWE-Bench, WMDP) and (2) verified company RSP data from official system cards (Anthropic ASL classifications, CBRN uplift scores). We map these to thresholds derived from international governance frameworks. Our key innovation is user-configurable thresholds with transparent justification labeling, distinguishing between data-driven thresholds (e.g., EU AI Act's 10^{25} FLOP) and our interpretations of qualitative frameworks (e.g., IDAIS Red Lines). The dashboard tracks 4 active capability metrics with configurable thresholds (WMDP, METR, SWE-Bench, Eval Awareness), 2 categorical metrics from company RSPs (ASL Classification, CBRN Uplift), and documents 5 inactive metrics awaiting data availability. Key findings include Claude Opus 4 being the first model released under ASL-3 Standard due to 2.53x CBRN bio uplift exceeding Anthropic's pre-registered 2x threshold. We document all sources in a comprehensive verification document and acknowledge significant data gaps where metrics lack published scores for most models. This tool aims to provide a shared reference point for discussing capability levels relative to governance concerns.

1. Introduction

AI capabilities are advancing rapidly. METR research shows task horizons doubling every 7 months, with Claude Opus 4.5 achieving 4h 49m autonomous operation (METR, Dec 2025). Training compute doubles every 6 months (Epoch AI). Multiple governance frameworks have emerged to define safety thresholds: the IDAIS Beijing consensus, EU AI Act, Seoul Frontier AI Commitments, and company Responsible Scaling Policies.

However, these frameworks face a critical challenge: they define concerns qualitatively while capability assessments are quantitative. There is no unified tool bridging this gap with transparent methodology.

Our main contributions are:

1. **A capability dashboard** aggregating 11 metrics from verified sources (METR, SWE-Bench, WMDP, UK AISI, Epoch AI)
2. **Framework-mapped thresholds** with explicit justification types (Data-Driven, Qualitative, Interpretation)

3. **User-configurable risk exploration** acknowledging that threshold choices involve value judgments
4. **Comprehensive source documentation** enabling independent verification of all claims

2. Related Work

Several efforts track AI capabilities or governance frameworks:

- **METR** publishes autonomous task horizon evaluations for select frontier models (metr.org).
- **SWE-Bench** provides public leaderboards for coding agent performance (swebench.com).
- **Epoch AI** estimates training compute for 3,200+ models (epoch.ai).
- **FLI AI Safety Index** rates companies on safety practices, not measured capabilities.
- **UK AISI Gray Swan** tested 22 models for adversarial robustness with aggregate findings.

Our work aggregates metrics from these sources and provides transparent threshold derivations linked to governance frameworks.

3. Methods

3.1 Data Sources (Verified)

We track capability metrics from two categories of sources:

Metric	Source	Coverage
Task Horizon	METR (metr.org)	~10 models
SWE-Bench	swebench.com	Public leaderboard
WMDP	Li et al. 2024 (ICML)	3 models
ASR	UK AISI Gray Swan	Aggregate only
Eval Awareness	arXiv:2505.23836	~5 models
ASL Classification	Anthropic System Cards	Claude models
CBRN Uplift	Anthropic System Cards	Claude models
FLI Safety Index	futureoflife.org	7 companies

3.2 Threshold Justification Types

We explicitly label each threshold by justification type:

- **Data-Driven:** Explicitly stated in source (only EU AI Act 10^{25} FLOP)
- **Qualitative:** Based on expert judgment without numeric specification
- **Interpretation:** Our translation of framework principles (e.g., all IDAIS thresholds)

3.3 Distance Calculation

$$\text{percentage_used} = (\text{current_value} / \text{threshold}) \times 100$$

Status: Safe (<65%), Warning (65-85%), Critical (85-100%), Breached (>100%). These thresholds are user-configurable.

4. Results

Key verified findings from published sources:

- **Anthropic RSP:** Claude Opus 4 was the first model released under ASL-3 Standard (May 2025) due to 2.53x CBRN bio uplift exceeding the pre-registered 2x threshold.
- **METR Task Horizon:** Claude Opus 4.5 achieves 4h 49m (95% CI: 1h 49m to 20h 25m), highest published. Task horizons doubling every 7 months.
- **SWE-Bench:** Claude Opus 4.5 agent system achieves 80.9% (note: SWE-Bench measures agent performance including scaffolding, not raw model capability).
- **Eval Awareness:** Gemini-2.5-Pro achieves 0.83 AUC, below human baseline of 0.92 (arXiv:2505.23836).
- **UK AISI Gray Swan:** 12.7% average ASR across 22 models; every model was vulnerable.
- **FLI Safety Index:** All companies scored D or F on Existential Safety (Winter 2025).

Important caveat: Most models in our dataset have data for only 1-2 metrics. Coverage is sparse.

5. Discussion and Limitations

Critical Limitations

- **Threshold interpretation:** All thresholds except EU AI Act compute are interpretations, not official specifications. IDAIS and Seoul provide no numeric thresholds.
- **Data sparsity:** We have verified data for only 12 models, most with 1-2 metrics each.
- **Proxy validity:** WMDP tests knowledge, not operational capability. SWE-Bench measures agents, not models.
- **Projection uncertainty:** Exponential projections assume constant doubling times, which may not hold.
- **Benchmark limitations:** Research (Apple's 'Illusion of Thinking', 2025) shows LLMs fail on high-complexity reasoning despite benchmark performance.

What This Tool Does NOT Do

- Provide official thresholds from any governance framework (except EU compute)
- Measure actual harm or risk (only capability proxies)
- Replace expert judgment on threshold appropriateness

6. Conclusion

The AI Safety Threshold Tracker demonstrates that connecting capability metrics to governance frameworks is feasible, though significant challenges remain. By making threshold assumptions

explicit, labeling justification types, and documenting all sources, we aim to enable more transparent discussions about capability levels. The tool's primary value is not in the specific thresholds (which are largely interpretations) but in providing a framework for exploring how different threshold assumptions lead to different conclusions about system safety.

Code and Data

- **Code repository:** <https://github.com/revathi-prasad/ai-safety-threshold-tracker>
- **Data files:** Curated JSON with inline source citations

References

1. Li, N. et al. (2024). The WMDP Benchmark. ICML 2024. arXiv:2403.03218
2. Needham, J. et al. (2025). Large Language Models Often Know When They Are Being Evaluated. arXiv:2505.23836
3. METR (2025). Claude Opus 4.5 Evaluation. metr.org
4. UK AISI/Gray Swan (2025). Agent Red-Teaming Challenge Results. grayswan.ai
5. Future of Life Institute (2025). AI Safety Index Winter 2025. futureoflife.org
6. European Parliament (2024). EU AI Act. artificialintelligenceact.eu
7. IDAIS (2024). Beijing Statement. idaais.ai
8. Epoch AI (2025). AI Model Database. epoch.ai
9. Anthropic (2025). Claude 4 System Card, Claude Opus 4.5 System Card. anthropic.com
10. Shojaei, P. et al. (2025). The Illusion of Thinking. arXiv:2506.06941

LLM Usage Statement

This project was developed with Claude (Anthropic). The human author provided project direction, platform design, and metric selection. Claude assisted with: literature review, code implementation, threshold interpretations, and report drafting. All data claims were verified against primary sources.

Appendix: Limitations and Dual-Use Considerations

Limitations

- **Data sparsity:** Most models have data for only 1-2 metrics. WMDP data is only available for 3 models from the original 2024 paper.
- **Threshold interpretation:** All thresholds except EU AI Act 10^{25} FLOP are LLM-generated interpretations of qualitative framework language. IDAIS and Seoul provide no numeric thresholds.
- **Proxy validity:** Benchmarks measure proxies, not actual risk. WMDP tests knowledge recall, not operational capability.
- **Projection uncertainty:** Exponential projections assume constant doubling times, which historically have not held.

Dual-Use Considerations

- **Capability signaling:** Aggregating capability data could help bad actors identify the most capable models for misuse. Mitigation: We only include publicly available data.
- **Threshold gaming:** Adjustable thresholds could be used to rationalize unsafe deployment. Mitigation: We label all thresholds as interpretations and encourage conservative defaults.

Ethical Considerations

This tool aims to support informed policy discussions, not replace expert judgment. Threshold choices involve value judgments that should be made by policymakers and domain experts, not automated systems.

Future Improvements

- Cross-framework threshold harmonization: Create mappings between ASL (Anthropic), CCL (DeepMind), and OpenAI Preparedness levels to enable cross-lab comparability, addressing the operationalization challenges identified in current RSP implementations.
- Expert-calibrated thresholds via structured elicitation: Develop consensus-based thresholds through structured elicitation processes (e.g., calibrated expert judgment, multi-stakeholder deliberation, prediction markets) with AI safety researchers, replacing our current LLM-generated interpretations with expert-validated values informed by risk tolerances in analogous domains (e.g., biosafety levels).
- Integration with third-party evaluation infrastructure: Connect to METR, UK AISI, and benchmark leaderboard APIs for real-time data feeds, supporting the third-party audit processes that governance frameworks increasingly require.
- Uncertainty quantification: Add confidence intervals to threshold proximity estimates and trajectory projections, acknowledging the measurement uncertainty inherent in capability evaluations.