FUTURE OF LIFE INSTITUTE | December 2025

AI Safety Index Winter 2025

Distinguished experts evaluate safety practices of leading AI companies across critical domains.

Full report at: **futureoflife.org/index** | Contact us: **policy@futureoflife.org**

Full Report

	Anthropic	OpenAI	Google Deepmind	xAI	Z.ai	Meta	DeepSeek	Alibaba Cloud
Overall Grade	C+	C+	C	D	D	D	D	D-
Score	2.67	2.31	2.08	1.17	1.12	1.10	1.02	0.98
Domains								
Risk Assessment (6 indicators)	B	B	C+	D	D+	D	D	D
Current Harms (7 indicators)	C+	C-	C	F	D	D+	D+	D+
Safety Frameworks (4 indicators)	C+	C+	C+	D+	D-	D+	F	F
Existential Safety (4 indicators)	D	D	D	F	F	F	F	F
Governance & Accountability (4 indicators)	B-	C+	C-	D	D	D	D	D+
Information Sharing (10 indicators)	A-	B	C	C	C-	D-	C-	D+
Survey Responses	✓	✓	✓	✓	✓	✗	✗	✗

Grading: Uses the US GPA system for grade boundaries: A+, A, A-, B+, [...], F letter values corresponding to numerical values 4.3, 4.0, 3.7, 3.3, [...], 0.

Executive Summary

- **A clear divide persists between the top performers (Anthropic, OpenAI, and Google DeepMind) and the rest of the companies reviewed (Z.ai, xAI, Meta, Alibaba Cloud, DeepSeek).** The most substantial gaps exist in the domains of risk assessment, safety framework, and information sharing, caused by limited disclosure, weak evidence of systematic safety processes, and uneven adoption of robust evaluation practices.
- **Existential safety remains the industry's core structural weakness.** All of the companies reviewed are racing toward AGI/superintelligence without presenting any explicit plans for controlling or aligning such smarter-than-human technology, thus leaving the most consequential risks effectively unaddressed.
- **Despite public commitments, companies' safety practices continue to fall short of emerging global standards.** While many companies partially align with these emerging standards, the depth, specificity, and quality of implementation remain uneven, resulting in safety practices that do not yet meet the rigor, measurability, or transparency envisioned by frameworks such as the EU AI Code of Practice.

About the Organization: The Future of Life Institute (FLI) is an independent nonprofit organization with the goal of reducing large-scale risks and steering transformative technologies to benefit humanity, with a particular focus on artificial intelligence (AI). Learn more at futureoflife.org.

What is the AI Safety Index?

Frontier AI systems are advancing rapidly, raising increasingly urgent questions about current harms and long-term controllability as models grow more autonomous, capable, and potentially self-improving. As capabilities grow, both the opportunities offered by these systems and the risks they pose expand accordingly. The AI Safety Index, developed by the Future of Life Institute with an independent panel of technical and governance experts, provides an impartial evaluation of how responsibly leading AI companies are approaching these challenges. Competitive pressures often reward profits over safety, so the Index aims to counterbalance those incentives by making companies' safety practices visible and comparable, creating reputational pressure to meet higher standards.

Methodology

The 2025 Winter AI Safety Index assesses the safety practices of eight leading frontier-model developers – Anthropic, Alibaba Cloud, DeepSeek, Google DeepMind, Meta, OpenAI, xAI, and Z.ai – across six critical domains, to foster transparency, promote robust safety practices, highlight areas for improvement and empower policymakers and the public to discern genuine safety measures from empty commitments.

An independent review panel of eight leading experts on technical and governance aspects of general-purpose AI volunteered to assess the companies' performances across 35 indicators of responsible conduct, contributing letter grades, brief justifications, and recommendations for improvement. The evaluation was supported by a comprehensive evidence base with company-specific information sourced from 1) publicly available material, including related research papers, policy documents, news articles, and industry reports, and 2) a tailored industry survey which firms could use to increase transparency around safety-related practices, processes and structures. The full list of indicators and collected evidence is presented in the full report.

Independent Review Panel

David Krueger is an Assistant Professor in Robust, Reasoning and Responsible AI in the Department of Computer Science and Operations Research (DIRO) at University of Montreal, a Core Academic Member at Mila, and an affiliated researcher at UC Berkeley's Center for Human-Compatible AI, and the Center for the Study of Existential Risk. His work focuses on reducing the risk of human extinction from AI.

Jessica Newman is the Founding Director of the AI Security Initiative, housed at the Center for Long-Term Cybersecurity at the University of California, Berkeley. She serves as an expert in the OECD Expert Group on AI Risk and Accountability and contributes to working groups within the U.S. Center for AI Standards and Innovation, EU Code of Practice Plenaries, and other AI standards and governance bodies.

Stuart Russell is a Professor of Computer Science at the University of California at Berkeley and Director of the Center for Human-Compatible AI and the Kavli Center for Ethics, Science, and the Public. He is a member of the National Academy of Engineering and a Fellow of the Royal Society. He is a recipient of the IJCAI Computers and Thought Award, the IJCAI Research Excellence Award, and the ACM Allen Newell Award. In 2021 he received the OBE from Her Majesty Queen Elizabeth and gave the BBC Reith Lectures. He coauthored the standard textbook for AI, which is used in over 1500 universities in 135 countries.

Tegan Maharaj is an Assistant Professor in the Department of Decision Sciences at HEC Montréal, where she leads the ERRATA lab on Ecological Risk and Responsible AI. She is also a core academic member at Mila. Her research focuses on advancing the science and techniques of responsible AI development. Previously, she served as an Assistant Professor of Machine Learning at the University of Toronto.

Dylan Hadfield-Menell is an Assistant Professor at MIT, where he leads the Algorithmic Alignment Group at the Computer Science and Artificial Intelligence Laboratory (CSAIL). A Schmidt Sciences AI2050 Early Career Fellow, his research focuses on safe and trustworthy AI deployment, with particular emphasis on multi-agent systems, human-AI teams, and societal oversight of machine learning.

Sharon Li is an Associate Professor in the Department of Computer Sciences at the University of Wisconsin-Madison. Her research focuses on algorithmic and theoretical foundations of safe and reliable AI, addressing challenges in both model development and deployment in the open world. She serves as the Program Chair for ICML 2026. Her awards include a Sloan Fellowship (2025), NSF CAREER Award (2023), MIT Innovators Under 35 Award (2023), Forbes 30under30 in Science (2020), and "Innovator of the Year 2023" (MIT Technology Review). She won the Outstanding Paper Award at NeurIPS 2022 and ICLR 2022.

Sneha Revanur is the founder and president of Encode, a global youth-led organization advocating for the ethical regulation of AI. Under her leadership, Encode has mobilized thousands of young people to address challenges like algorithmic bias and AI accountability. She was featured on TIME's inaugural list of the 100 most influential people in AI.

Yi Zeng is an AI Professor at the Chinese Academy of Sciences, the Founding Dean of the Beijing Institute of AI Safety and Governance, and the Director of the Beijing Key Laboratory of Safe AI and Superalignment. He serves on the UN High-level Advisory Body on AI, the UNESCO Ad Hoc Expert Group on AI Ethics, the WHO Expert Group on the Ethics/Governance of AI for Health, and the National Governance Committee of Next Generation AI in China. He has been recognized by the TIME100 AI list.