July 2025

FLI AI Safety Index: Summer 2025

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

Full report at: **futureoflife.org/index-s25** | Contact us: **policy@futureoflife.org**

Full Report

Domains Number of indicators	Anthropic	OpenAI	Google Deepmind	x.AI	Meta	Zhipu AI	Deepseek
Overall Grade	C+	C	C-	D	D	F	F
Overall Score	2.64	2.10	1.76	1.23	1.06	0.62	0.37
Risk Assessment 6 indicators	C+	C	C-	F	D	F	F
Current Harms 8 indicators	B-	B	C+	D+	D+	D	D-
Safety Frameworks 6 indicators	C	C	D+	D+	D+	F	F
Existential Safety 5 indicators	D	F	D-	F	F	F	F
Governance & Accountability 5 indicators	A-	C-	D	C-	D-	D+	D+
Information Sharing 4 indicators	A-	B	B	C+	D	D	F

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

The Summer 2025 FLI AI Safety Index revealed modest progress by some companies in certain domains, but a striking lack of commitment to many areas of safety, most notably control loss risk.

- **Large disparities remain in risk management practices:** Some companies have expanded their safety frameworks in promising ways and conducted serious risk assessment efforts, but others still fail to implement even the most basic precautions. None have robust, reliable, plans for ensuring the public remains safe from their products.
- **The Control-Problem remains unsolved:** Despite their explicit ambitions to develop artificial general intelligence (AGI) capable of rivaling or exceeding human intelligence and replacing humans, the review panel again found the current strategies of all companies inadequate for ensuring that these systems remain safe and under human control, with only minor efforts to solve this problem.
- **External oversight is necessary:** The report showed how companies had abandoned previous voluntary commitments, highlighting a need for third party validation of risk assessments and legally binding oversight with enforcement mechanisms.

Context

AI systems are growing increasingly powerful as tech companies drive toward artificial general intelligence (AGI) and beyond. Just as functioning breaks give drivers the confidence to accelerate, effective AI safety measures give society the confidence to innovate and adopt AI. Competitive pressures can incentivize a race to the bottom that prioritizes profits over safety, so to improve incentives, the Future of Life Institute periodically convenes an independent panel of leading AI experts to conduct a comprehensive safety review of prominent tech companies.

Methodology

The 2025 FLI AI Safety Index evaluates the safety practices of seven leading general-purpose AI (GPAI) companies—Anthropic, OpenAI, Google DeepMind, x.AI, Meta, Zhipu AI, and Deepseek—across six critical domains, to foster transparency, promote robust safety practices, highlight areas for improvement and empower the public to discern genuine safety measures from empty commitments.

An independent review panel of leading experts on technical and governance aspects of general-purpose AI volunteered to assess the companies' performances across 34 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

Dylan Hadfield-Menell is the Bonnie and Marty (1964) Tenenbaum Career Development 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.

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 also serves as the Director of the UC Berkeley AI Policy Hub, an expert in the OECD Expert Group on AI Risk and Accountability, and a member of the U.S. AI Safety Institute Consortium.

Stuart Russell is a Professor of Computer Science at the University of California at Berkeley, holder of the Smith-Zadeh Chair in Engineering, and Director of the Center for Human-Compatible AI and the Kavli Center for Ethics, Science, and the Public. 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 co-authored 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.

Sneha Revanur is the founder and president of Encode Justice, a global youth-led organization advocating for the ethical regulation of AI. Under her leadership, Encode Justice 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.

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, and a Core Academic Member at Mila, 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 artificial intelligence through technical research as well as education, outreach, governance and advocacy.

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.