Future of Life Institute

AI Safety Index

Summer 2025

17th July 2025

Available online at: futureoflife.org/index
Contact us: policy@futureoflife.org

future of life INSTITUTE

Contents

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.

1 Executive Summary

The Future of Life Institute's AI Safety Index provides an independent assessment of seven leading AI companies' efforts to manage both immediate harms and catastrophic risks from advanced AI systems. Conducted with an expert review panel of distinguished AI researchers and governance specialists, this second evaluation reveals an industry struggling to keep pace with its own rapid capability advances—with critical gaps in risk management and safety planning that threaten our ability to control increasingly powerful AI systems.

	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	C+	C	C-	F	D	F	F
Current Harms	B-	B	C+	D+	D+	D	D
Safety Frameworks	C	C	D+	D+	D+	F	F
Existential Safety	D	F	D-	F	F	F	F
Governance & Accountability	A-	C-	D	C-	D-	D+	D+
Information Sharing	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.

1.1 Key Findings

- **Anthropic gets the best overall grade (C+).** The firm led on risk assessments, conducting the only human participant bio-risk trials, excelled in privacy by not training on user data, conducted world-leading alignment research, delivered strong safety benchmark performance, and demonstrated governance commitment through its Public Benefit Corporation structure and proactive risk communication.
- **OpenAI secured second place ahead of Google DeepMind.** OpenAI distinguished itself as the only company to publish its whistleblowing policy, outlined a more robust risk management approach in its safety framework, and assessed risks on pre-mitigation models. The company also shared more details on external model evaluations, provided a detailed model specification, regularly disclosed instances of malicious misuse, and engaged comprehensively with the AI Safety Index survey.
- **The industry is fundamentally unprepared for its own stated goals.** Companies claim they will achieve artificial general intelligence (AGI) within the decade, yet none scored above D in Existential Safety planning. One reviewer called this disconnect "deeply disturbing," noting that despite racing toward human-level AI, "none of the companies has anything like a coherent, actionable plan" for ensuring such systems remain safe and controllable.
- **Only 3 of 7 firms report substantive testing for dangerous capabilities linked to large-scale risks such as bio- or cyber-terrorism** (Anthropic, OpenAI, and Google DeepMind). While these leaders marginally improved the quality of their model cards, one reviewer warns that the underlying safety tests still miss basic risk-assessment standards: "The methodology/reasoning explicitly linking a given evaluation or experimental procedure to the risk, with limitations and qualifications, is usually absent. [...] I have very