11th December 2024

FLI AI Safety Index 2024

Rapidly improving AI capabilities have increased interest in how companies report, assess and attempt to mitigate associated risks. The 2024 FLI AI Safety Index therefore convened an independent panel of seven distinguished AI and governance experts to evaluate the safety practices of six leading general-purpose AI companies across six critical domains.

View Full Report

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

Firm	Overall Grade	Score	Risk Assessment	Current Harms	Safety Frameworks	Existential Safety Strategy	Governance & Accountability	Transparency & Communication
Anthropic	C	2.13	C+	B-	D+	D+	C+	D+
Google DeepMind	D+	1.55	C	C+	D-	D	D+	D
OpenAI	D+	1.32	C	D+	D-	D-	D+	D-
Zhipu AI	D	1.11	D+	D+	F	F	D	C
x.AI	D-	0.75	F	D	F	F	F	C
Meta	F	0.65	D+	D	F	F	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.

Key Findings

- **Large risk management disparities:** While some companies have established initial safety frameworks or conducted some serious risk assessment efforts, others have yet to take even the most basic precautions.
- **Jailbreaks:** All the flagship models were found to be vulnerable to adversarial attacks.
- **Control-Problem:** Despite their explicit ambitions to develop artificial general intelligence (AGI), capable of rivaling or exceeding human intelligence, the review panel deemed the current strategies of all companies inadequate for ensuring that these systems remain safe and under human control.
- **External oversight:** Reviewers consistently highlighted how companies were unable to resist profit-driven incentives to cut corners on safety in the absence of independent oversight. While Anthropic's current and OpenAI's initial governance structures were highlighted as promising, experts called for third-party validation of risk assessment and safety framework compliance across all companies.

Methodology

The Index aims to foster transparency, promote robust safety practices, highlight areas for improvement and empower the public to discern genuine safety measures from empty claims.

An independent review panel of leading experts on technical and governance aspects of general-purpose AI volunteered to assess the companies' performances across 42 indicators of responsible conduct, contributing letter grades, brief justifications, and recommendations for improvement. The panellist selection focused on academia rather than industry to reduce potential conflicts of interest.

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 attached to the report.

Independent Review Panel

Yoshua Bengio is a Full Professor in the Department of Computer Science and Operations Research at Université de Montreal, as well as the Founder and Scientific Director of Mila and the Scientific Director of IVADO. He is the recipient of the 2018 A.M. Turing Award.

Atoosa Kasirzadeh is an Assistant Professor at Carnegie Mellon University. Previously, she was a visiting faculty researcher at Google, a Chancellor's Fellow and Director of Research at the Centre for Technomoral Futures at the University of Edinburgh.

David Krueger is an Assistant Professor 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.

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.

Sneha Revanur is the founder and president of Encode Justice, a global youth-led organization advocating for the ethical regulation of AI. TIME featured her as one of the 100 most influential people in AI.

Jessica Newman is the Director of the AI Security Initiative (AISI), housed at the UC Berkeley Center for Long-Term Cybersecurity. She is also a Co-Director of the UC Berkeley AI Policy Hub.

Stuart Russell OBE is a Professor of Computer Science at UC 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 co-authored the standard textbook for AI, which is used in over 1500 universities in 135 countries.

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.