



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

1 Executive Summary	2
1.1 Key Findings	2
1.2 Improvement opportunities by company	3
1.3 Methodology	4
1.4 Independent review panel	5
2 Introduction	6
3 Methodology	7
3.1 Companies Assessed	7
3.2 Index Design and Structure	7
3.3 Related Work and Incorporated Research	10
3.4 Data Sources and Evidence Collection	10
3.5 Grading Process and Expert Review	11
3.6 Limitations	11
4 Results	13
4.1 Key Findings	13
4.2 Improvement opportunities by company	14
4.3 Domain-level findings	15
5 Conclusions	20
Appendix A: Grading Sheets	21
Risk Assessment	22
Current Harms	33
Safety Frameworks	41
Existential Safety	48
Governance & Accountability	59
Information Sharing	71
Appendix B: Company Survey	85
Introduction	85
Whistleblowing Policies (15 Questions)	86
External Pre-Deployment Safety Testing (6 Questions)	91
Internal Deployments (3 Questions)	94
Safety Practices, Frameworks, and Teams (9 Questions)	95

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](https://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

low confidence that dangerous capabilities are being detected in time to prevent significant harm. Minimal overall investment in external 3rd party evaluations decreases my confidence further.”

- **Capabilities are accelerating faster than risk management practice**, and the gap between firms is widening. With no common regulatory floor, a few motivated companies adopt stronger controls while others neglect basic safeguards, highlighting the inadequacy of voluntary pledges.
- **Whistleblowing policy transparency remains a weak spot.** Public whistleblowing policies are a common best practice in safety-critical industries because they enable external scrutiny. Yet, among the assessed companies, only OpenAI has published its full policy, and it did so only after media reports revealed the policy’s highly restrictive non-disparagement clauses.
- **Chinese AI firms Zhipu.AI and DeepSeek both received failing overall grades.** However, the report scores companies on norms such as self-governance and information-sharing, which are far less prominent in Chinese corporate culture. Furthermore, as China already has regulations for advanced AI development, there is less reliance on AI safety self-governance. This is in contrast to the United States and United Kingdom, where the other companies are based, and which have, as yet, passed no such regulation on frontier AI.
 - *Note: the scoring was completed in early July and does not reflect recent events such as xAI’s Grok 4 release, Meta’s superintelligence announcement, or OpenAI’s commitment to sign the EU AI Act Code of Practice.*

These findings reveal an unregulated industry where competitive pressures and technological ambition far outpace safety infrastructure and norms. This imbalance becomes more dangerous as companies pursue their stated goal of achieving artificial general intelligence (AGI) that matches or exceeds human capabilities within the decade.

1.2 Improvement opportunities by company

Here we highlight examples of how individual companies can improve future scores with relatively modest effort.

Anthropic:

- Publish a full whistleblowing policy to match OpenAI’s transparency standard.
- Become more transparent and explicit about risk assessment methodology—e.g. why/how exactly is the particular eval related to a (class of) risks. Include reasoning in model cards that explicitly links evaluations or experimental procedures to specific risk, with limitations and qualifications.

OpenAI:

- Rebuild lost safety team capacity and demonstrate renewed commitment to OpenAI’s original mission.
- Maintain the strength of current non-profit governance elements to guard against financial pressures undermining OpenAI’s mission.

Google DeepMind:

- Publish a full whistleblowing policy to match OpenAI’s transparency standard.
- Publish evaluation results for models without safety guardrails to more closely approximate true model capabilities.
- Improve coordination between DeepMind safety team and Google’s policy team.
- Increase transparency around and investment in third-party model evaluations for dangerous capabilities.

xAI:

- Ramp up risk assessment efforts and publish implemented evaluations in upcoming model cards.
- Boost current draft safety framework to match the efforts by Anthropic and OpenAI.
- Publish a full whistleblowing policy to match OpenAI’s transparency standard.

Meta:

- Significantly increase investment in technical safety research, especially tamper-resistant safeguards for open-weight models.
- Ramp up risk assessment efforts and publish implemented evaluations in upcoming model cards.
- Publish a full whistleblowing policy to match OpenAI’s transparency standard.

Zhipu AI:

- Publish the AI Safety Framework promised at the AI Summit in Seoul.
- Ramp up risk assessment efforts and publish implemented evaluations in upcoming model cards.

DeepSeek:

- Address extreme jailbreak vulnerability before next release.
- Ramp up risk assessment efforts and publish implemented evaluations in upcoming model cards.
- Develop and publish a comprehensive AI safety framework.

All companies: Publish a first concrete plan, however imperfect, for how they hope to control the AGI/ASI they plan to build.

1.3 Methodology

Index Structure: The 2025 Index evaluates seven leading AI companies on 33 indicators spanning six critical domains.

Risk Assessment

Internal Testing

- Dangerous Capability Evaluations
- Elicitation for Dangerous Capability Evaluations
- Human Uplift Trials

External Testing

- Independent Review of Safety Evaluations
- Pre-deployment External Safety Testing
- Bug Bounties for Model Vulnerabilities

Current Harms

Model Safety / Trustworthiness

- Stanford's HELM Safety Benchmark
- Stanford's HELM AIR Benchmark
- TrustLLM Benchmark

Robustness

- Gray Swan Arena: UK AISI Agent Red-Teaming Challenge
- Cisco Security Risk Evaluation
- Protecting Safeguards from Fine-tuning

Digital Responsibility

- Watermarking
- User Privacy

Safety Frameworks

Risk Identification

Risk Analysis and Evaluation

Risk Treatment

Risk Governance

Existential Safety

Existential Safety Strategy

Internal Monitoring and Control Interventions

Technical AI Safety Research

Supporting External Safety Research

Governance & Accountability

Lobbying on AI Safety Regulations

Company Structure & Mandate

Whistleblowing

- Whistleblowing Policy Transparency
- Whistleblowing Policy Quality Analysis
- Reporting Culture & Whistleblowing Track Record

Information Sharing

Technical Specifications

- System Prompt Transparency
- Behavior Specification Transparency

Voluntary Cooperation

- G7 Hiroshima AI Process Reporting
- FLI AI Safety Index Survey Engagement

Risks & Incidents

- Serious Incident Reporting & Government Notifications
- Extreme-Risk Transparency & Engagement

Data Collection: Evidence was gathered between March 24 and June 24, 2025, combining publicly available materials—including model cards, research papers, and benchmark results—with responses from a targeted company survey designed to address specific transparency gaps in the industry, such as transparency on whistleblower protections and external model evaluations. The complete evidence base is documented in [Appendix A](#) and [Appendix B](#).

Expert Evaluation: An independent panel of leading AI researchers and governance experts reviewed company-specific evidence and assigned domain-level grades (A-F) based on absolute performance standards. Reviewers provided written justifications and improvement recommendations. Final scores represent averaged expert assessments, with individual grades kept confidential.

1.4 Independent review panel

The scoring was conducted by a panel of distinguished AI experts:



Dylan Hadfield-Menell

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.



Tegan Maharaj

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.



Jessica Newman

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.



Sneha Revanur

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.



Stuart Russell

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 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 co-authored the standard textbook for AI, which is used in over 1500 universities in 135 countries.



David Krueger

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 artificial intelligence through technical research as well as education, outreach, governance and advocacy.

2 Introduction

Artificial intelligence systems are becoming more capable and autonomous at an unprecedented pace. Fueled by massive investments and technical breakthroughs, these general-purpose AI systems are transforming from specialized tools into increasingly versatile agents, being deployed in increasingly high-stakes settings. These trends pose significant risks, ranging from **malicious use to systemic failures and loss of meaningful human control**. This makes independent scrutiny of how GPAI systems are developed and deployed more urgent than ever.

The AI Safety Index is a response to that urgency. Developed and published by the Future of Life Institute (FLI), in collaboration with a review panel composed of some of the foremost independent experts in AI, the Index provides an independent assessment of how responsibly the world's leading AI companies are developing and deploying increasingly powerful AI systems. It focuses on institutional safeguards, such as risk management frameworks, third-party oversight, and whistleblower policies. Each company is evaluated on a set of 33 indicators across six domains, and scores are presented in a format designed to be accessible to both expert and non-expert audiences.

Since the release of the [inaugural Index](#) in December 2024, the global awareness around AI risks has continued to increase. Mandated by the nations attending the AI Safety Summit in the UK, the first [International AI Safety Report](#) synthesized the current scientific understanding of AI risks and potential mitigation strategies, acknowledging the potential for catastrophic outcomes. Distinguished subject experts from industry, government, and academia convened in Singapore and agreed on the [Singapore Consensus](#) on Global AI Safety Research Priorities, which outlines key research priorities for addressing risks from advanced AI. These developments confirm a growing international consensus: keeping AI safe as capabilities advance demands urgent investment in alignment research and substantial improvements in risk management.

The Summer 2025 edition of the FLI AI Safety Index evaluates seven frontier AI companies—OpenAI, Anthropic, Google DeepMind, Meta, xAI, Zhipu AI, and, for the first time, DeepSeek—using updated and improved indicators that reflect evolving deployment practices and safety norms. The Index is intended to be a practical, public-facing tool for tracking corporate behavior, identifying emerging best practices, and surfacing critical gaps. By making companies' risk management practices more visible and comparable, the Index aims to strengthen incentives for responsible AI development and to close the gap between safety commitments and real-world actions.

3 Methodology

Evaluating safety practices at the cutting edge of AI development requires a flexible and evolving methodology that can capture and appropriately weigh both concrete implementations and stated positions, commitments, and levels of transparency across diverse organizations.

This section outlines how the AI Safety Index evaluates and grades AI companies. We explain indicator design and structure, our evidence collection and data sources, the independent review process, and discuss notable limitations. By making our methods fully transparent, we aim to provide stakeholders with the context required to understand both the strengths and limitations of our assessments.

3.1 Companies Assessed

The 2025 FLI AI Safety Index evaluates seven leading general-purpose AI developers: Anthropic, OpenAI, Google DeepMind, Meta, Zhipu AI, x.AI, and DeepSeek. These companies represent the global frontier of AI capability development and were selected based on flagship model performance. The inclusion of the Chinese firms Zhipu AI and DeepSeek also reflects our intention to make the Index representative of leading developers globally. The inclusion of DeepSeek for this second iteration recognizes both its technical achievements in efficient model training and its growing influence in the Chinese AI ecosystem. Future iterations of the Index may assess different companies as the competitive landscape evolves.

Our selection focused on firms that have deployed models with competitive performance on public benchmarks. Therefore, we excluded companies that aim to build artificial general intelligence but have not yet deployed frontier models, such as [Safe Superintelligence Inc.](#)

3.2 Index Design and Structure

The AI Safety Index evaluates companies on a set of 33 indicators across six domains that capture different aspects of responsible AI development and deployment. Each domain contains multiple indicators along which differences in responsible AI practices between firms become visible. The indicators were selected based on the five criteria below.

- **High signal value:** Revealing substantive differences in companies' investments and priorities
- **Implementation focus:** Prioritizing demonstrated measures over stated commitments
- **Information availability:** Ensuring sufficient public evidence exists for evaluation
- **Clear definition:** Enabling consistent evaluation across companies
- **Leadership recognition:** Rewarding exceptional practices while maintaining adequate standards

Complete indicator definitions, rationales, and sources are provided in [Appendix A](#). Below, we briefly introduce each of the 33 indicators:

☰ Risk Assessment

This domain evaluates the rigor and comprehensiveness of companies' risk identification and assessment processes for their current flagship models. The focus is on implemented assessments, not stated commitments.

Group	Indicator Title	Summary
<i>Internal testing</i>	Dangerous Capability Evaluations	Tracks whether developers assess AI systems for harmful capabilities like cyber-offense, autonomous replication, or influence operations.
	Elicitation for Dangerous Capability Evaluations	Evaluates how transparently companies disclose and share their elicitation strategy used in dangerous capability evaluations.
	Human Uplift Trials	Evaluates whether companies conduct controlled experiments to measure how AI may increase users' ability to cause real-world harm.
<i>External testing</i>	Independent Review of Safety Evaluations	Assess whether third-party experts independently verify and critique the quality and accuracy of a developer's safety evaluations.
	Pre-Deployment External Safety Testing	Measures whether independent, unaffiliated experts are given meaningful access to test a model's safety before public release.
	Bug Bounties for Model Vulnerabilities	Assess whether developers offer structured incentives for discovering and disclosing safety issues specific to AI model behavior.

🏠 Current Harms

This domain covers demonstrated safety outcomes rather than commitments or processes. It focuses on the AI model's performance on safety benchmarks and the robustness of implemented safeguards against adversarial attacks.

<i>Model Safety / Trustworthiness</i>	Stanford's HELM Safety Benchmark	Evaluates how language models perform on key safety metrics like robustness, fairness, and resistance to harmful behavior.
	Stanford's HELM AIR Benchmark	Measures AI model safety and security on benchmark aligned with emerging government regulations and company policies.
	TrustLLM Benchmark	Assesses a model's trustworthiness across dimensions such as safety, ethics, and alignment with human values and expectations.
<i>Robustness</i>	Gray Swan Arena (UK AISI Agent Red-Teaming Challenge)	Tests how AI agents withstand adversarial challenges in high-risk, simulated decision-making environments to expose vulnerabilities.
	Cisco Security Risk Evaluation	Reports cybersecurity assessments of AI systems, focusing on their vulnerability to automated jailbreaking attacks.
	Protecting Safeguards from Fine-tuning	Evaluates whether AI providers implement protections that prevent fine-tuning from disabling important safety mechanisms or filters.
<i>Digital Responsibility</i>	Watermarking	Assess whether AI outputs are marked in a detectable way to help track origin and reduce misinformation or misuse.
	User Privacy	Measures the degree to which an AI company protects user data from extraction, exposure, or inappropriate use by models.

🔧 Safety Frameworks

This domain evaluates the companies' published safety frameworks for frontier AI development and deployment from a risk management perspective. This comprehensive analysis was conducted by the non-profit research organisation [SaferAI](#).

Risk Identification	Evaluates whether companies systematically identify AI risks through comprehensive methods, including literature review, red teaming, and diverse threat modeling techniques.
Risk Analysis and Evaluation	Assesses whether companies translate abstract risk tolerances into concrete, measurable thresholds that trigger specific responses.
Risk Treatment	Measures whether companies implement comprehensive mitigation strategies across containment, deployment safeguards, and affirmative safety assurance, with continuous monitoring throughout the AI lifecycle.
Risk Governance	Examines whether companies establish clear risk ownership, independent oversight, safety-oriented culture, and transparent disclosure of their risk management approaches and incidents.

Existential Safety

This domain examines companies' preparedness for managing extreme risks from future AI systems that could match or exceed human capabilities, including stated strategies and research for alignment and control.

Existential Safety Strategy	Assesses whether companies developing AGI publish credible, detailed strategies for mitigating catastrophic and existential AI risks, including alignment and control, governance, and planning.
Internal Monitoring and Control Interventions	Evaluates whether companies implement technical controls and protocols to detect and prevent model misalignment during internal use.
Technical AI Safety Research	Tracks whether companies publish research relevant to extreme-risk mitigation, including areas like interpretability, scalable oversight, and dangerous capability evaluations.
Supporting External Safety Research	Assesses the extent to which companies support independent AI safety work through mentorships, funding, model access, and collaboration with external researchers.

Governance & Accountability

This domain audits whether each company's governance structure and day-to-day operations prioritize meaningful accountability for the real-world impacts of its AI systems. Indicators examine whistleblowing systems, legal structures, and advocacy on AI regulations.

Lobbying on AI Safety Regulations		Tracks how a company has engaged in lobbying or public advocacy that influences laws and policies related to AI safety.
Company Structure & Mandate		Evaluates whether a company's legal and governance setup includes enforceable commitments that prioritize safety over profit incentives.
Whistleblowing Protections	Whistleblowing Policy Transparency	Assesses how publicly accessible and complete a company's whistleblowing system is, including reporting channels, protections, and transparency of outcomes.
	Whistleblowing Policy Quality Analysis	Rates the comprehensiveness and alignment of a company's whistleblowing policy with international best practices and AI-specific safety needs.
	Reporting Culture & Whistleblowing Track Record	Examines whether the company climate makes employees feel they can safely report AI safety concerns, based on leadership behavior, third-party evidence, and past incidents.

Information Sharing

This section gauges how openly firms share information about products, risks, and risk management practices. Indicators cover voluntary cooperation, transparency on technical specifications, and risk/incident communication.

Technical Specifications	System Prompt Transparency	Assesses whether companies publicly disclose the actual system prompts used in their deployed AI models, including version histories and design rationales.
	Behavior Specification Transparency	Evaluates if developers publish detailed and up-to-date documentation explaining their models' intended behavior, values, and decision-making logic across diverse scenarios.
Voluntary Cooperation	G7 Hiroshima AI Process Reporting	Tracks whether companies submitted detailed safety and governance disclosures to the G7 Hiroshima AI Process, reflecting their commitment to transparency.
	FLI AI Safety Index Survey Engagement	Reports that companies voluntarily completed and submitted FLI's detailed safety survey to supplement publicly available information.
Risks & Incidents	Serious Incident Reporting & Government Notifications	Evaluates public commitments, frameworks, and track records around reporting serious AI-related incidents to governments and peers.
	Extreme-Risk Transparency & Engagement	Measures whether company leaders publicly acknowledge catastrophic AI risks and proactively communicate those concerns to external audiences.

3.3 Related Work and Incorporated Research

Related Work: Several notable related efforts that drive transparency and accountability within the industry continue to inspire and complement the AI Safety Index. The most comprehensive of these efforts include [SaferAI's in-depth analysis](#) and ranking of AI companies' public safety frameworks, and two projects by Zach Stein-Perlman—[AILabWatch.org](#) and [AISafetyClaims.org](#)—which provide detailed, technical evaluations of how leading AI companies work to avert catastrophic risks from advanced AI.

Incorporated Research: Where appropriate, the 2025 Index incorporates existing comparative analysis led by credible research institutions. The Safety Framework domain imports SaferAI's [in-depth assessment](#) of firms' published safety frameworks in the 'Safety Framework' domain. [SaferAI](#) is a leading governance and research non-profit with significant expertise in AI risk management. The Index further incorporates [AILabWatch.org](#)'s tracker of technical AI safety research in the 'Existential Safety' domain. Our research on the quality of companies' whistleblowing policies in the 'Governance & Accountability' domain was enabled through support from [OASIS](#), a non-profit supporting individuals working at the frontier of AI who want to flag risks.

The 'Current Harms' domain evaluates flagship model performance on leading safety benchmarks, including the [TrustLLM](#) benchmark, and the [HELM AIR-Bench](#) and [HELM Safety](#) benchmarks by [Stanford's Center for Research on Foundation Models](#). The section further features results from the [UK AI Security Institute's Red-teaming Challenge](#) on the [Gray Swan](#) Arena, and a model security analysis from [Cisco](#).

3.4 Data Sources and Evidence Collection

The evidence collection process for the 2025 AI Safety Index was conducted between March 24 and June 24, 2025, using publicly available information and a dedicated company survey for additional voluntary disclosures. Throughout the data collection process, FLI aimed to minimize bias and ensure a fair evaluation by applying consistent search protocols and evidence standards across companies.

Desk research: Our primary evidence base consists of public documentation that companies have released about their AI systems and risk management practices. This includes technical model cards detailing capabilities and limitations, peer-reviewed research papers on safety methodologies, official policy documents, blog posts outlining safety commitments, and recordings or transcripts of leadership interviews or testimony before government bodies. We further incorporated metrics of flagship model performance on external safety benchmarks, news reports from credible media outlets, and reports of relevant assessments by independent research organizations.

Company survey: To supplement public information, FLI created a 34-question survey that addresses current gaps in voluntary disclosures. The survey was sent out via e-mail on May 28, and firms were given until June 17 to respond. The survey can be reviewed in full in [Appendix B](#). Compared to the previous winter 2024 iteration of the Index, the updated survey was shorter and more specifically targeted on risk management-related domains where current transparency standards in the industry are lacking, such as whistleblowing policies, external third-party model evaluations, and internal AI deployment practices. We received survey responses from three companies (OpenAI, Zhipu AI, and xAI), representing 43% of assessed firms. Anthropic, Google DeepMind, Meta, and DeepSeek have not submitted a response. Full survey responses are attached in [Appendix B](#) as well.

Grading Sheets: The resulting evidence base underlying the index was then structured into the grading sheets found in [Appendix A](#). The grading sheets, which are split into six domains, contain company-specific information for each of the 33 indicators of the current edition of the index. For each indicator, the grading

sheets present a definition of its scope, a rationale for the inclusion of the indicator, and references to relevant literature where appropriate. The sources for all pieces of company-specific information are embedded in the relevant locations with hyperlinks. We prioritized primary sources directly from companies over secondary reporting wherever possible, with investigative journalism providing valuable insights, uncovering practices not voluntarily disclosed. Where indicators overlap with survey questions, relevant survey responses were highlighted in the grading sheets.

3.5 Grading Process and Expert Review

The 2025 AI Safety Index's grading process was designed to ensure an impartial and qualified evaluation of the companies' performance across the selected indicators. It features a review panel of distinguished independent experts who assess the company-specific evidence for each indicator and assign domain-level grades that represent companies' performance within these domains.

Review Panel: To ensure that the Index scores rest upon authoritative judgements, FLI selected a group of six leading independent experts to grade company performance on the set of indicators. Panel members were selected for their domain expertise and absence of conflicts of interest. The panel's composition further reflects a diversity of backgrounds, given that the Index spans from technical AI Safety topics to the domains of governance and policy. The panel thus features both renowned machine learning professors who specialize in alignment and control, and also governance experts from the non-profit sector. The composition of the panel remained mostly unchanged from the previous version of the index. We are grateful to Dylan Hadfield-Menell for joining the panel as a new member, replacing Yoshua Bengio, who stepped back due to time constraints from competing professional commitments. Review panel member Atoosa Kasirzadeh had to pause her contribution for the current version due to a family emergency. The review panel is introduced at the beginning of this document.

Grading Phase: Grading sheets and survey results were shared with the review panel for evaluation on June 24, and the grading period ended on July 9. After reviewing the evidence, reviewers assigned letter grades (A+ to F) to each company per domain. For each grade, reviewers could provide brief justifications and recommendations. They were also able to provide domain-level comments when feedback applied to multiple firms or to explain their judgments. Not every reviewer graded every domain, but experts were assigned domains relevant to their area of expertise. Importantly, no fixed weighting was imposed across indicators within a domain. This approach allowed expert reviewers to apply their judgment in emphasizing aspects they deemed most critical. The grading sheets provided to reviewers further contained grading scales based on absolute performance standards rather than relative rankings, ensuring consistent expectations regardless of company size or geography. Final domain scores were calculated by averaging all reviewer grades for a given domain. Overall grades represent the average across all domains.

3.6 Limitations

Our methodology has several important limitations that should be considered when interpreting the Index results.

| Information Availability and Verification

Our evaluation relies primarily on public information, which creates fundamental constraints. Companies control their disclosure levels, making it difficult to distinguish between poor transparency and poor implementation. We designed indicators around these transparency constraints, focusing where meaningful differences between companies were identifiable. For example, we cannot assess critical practices such as cybersecurity investments to protect model weights, as this information is rarely disclosed publicly.

The 33 indicators represent a subset of important practices for which meaningful evidence exists, not a comprehensive assessment of all safety dimensions. Furthermore, we cannot independently verify company claims and must assume official reports are truthful—a significant limitation given the high stakes involved.

| Geographic and Cultural Context

Our methodology, developed in Western academic contexts, is rather Western-centric and may adversely impact the scores of the Chinese companies Zhipu and DeepSeek. For example, it places a premium on self-governance and information-sharing, both of which are far less prominent in Chinese corporate culture. As China already has regulations for generative AI in place, firms face less incentive to self-regulate when it comes to AI safety. This is in contrast to the US and UK, where the other companies are based, and where no such regulation exists.

Moreover, information availability varies dramatically across regions. U.S.-based companies sometimes provide extensive documentation, from detailed model cards to public safety frameworks. Chinese companies such as Zhipu AI and DeepSeek operate under different regulatory frameworks and cultural norms around transparency, making direct comparisons challenging. Language barriers compound these challenges, potentially affecting our assessment of non-English resources. Several indicators have limited applicability to Chinese firms operating in fundamentally different contexts.

| Methodological Constraints














Our focus on observable, documentable practices may undervalue crucial but hard-to-measure factors such as safety culture. Additionally, while our six-member panel brings diverse expertise, it cannot encompass all relevant domains. Reviewers' backgrounds inevitably influence assessments, and the flexibility in weighting indicators may introduce inconsistencies.

| Moving Forward

We aim to mitigate these limitations through rigorous documentation of sources, methodology, and reviewer materials. Readers should interpret Index results as one input among many for understanding AI safety practices. We invite constructive criticism and helpful suggestions at policy@futureoflife.org and are committed to improving the project with every iteration.

4 Results

Overall Rankings: Anthropic leads with a C+ (2.64), followed by OpenAI (C, 2.10) and Google DeepMind (C-, 1.76). The middle tier includes x.AI and Meta (both D), while Chinese companies Zhipu AI and DeepSeek trail with failing grades. Notably, no company achieved higher than C+, indicating that even industry leaders fall short of adequate safety standards.

	 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.

4.1 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 low confidence that dangerous capabilities are being detected in time to prevent significant harm. Minimal overall investment in external 3rd party evaluations decreases my confidence further.”

- **Capabilities are accelerating faster than risk management practice**, and the gap between firms is widening. With no common regulatory floor, a few motivated companies adopt stronger controls while others neglect basic safeguards, highlighting the inadequacy of voluntary pledges.
- **Whistleblowing policy transparency remains a weak spot.** Public whistleblowing policies are a common best practice in safety-critical industries because they enable external scrutiny. Yet, among the assessed companies, only OpenAI has published its full policy, and it did so only after media reports revealed the policy's highly restrictive non-disparagement clauses.
- **Chinese AI firms Zhipu.AI and DeepSeek both received failing overall grades.** However, the report scores companies on norms such as self-governance and information-sharing, which are far less prominent in Chinese corporate culture. Furthermore, as China already has regulations for advanced AI development, there is less reliance on AI safety self-governance. This is in contrast to the United States and United Kingdom, where the other companies are based, and which have, as yet, passed no such regulation on frontier AI.
 - *Note: the scoring was completed in early July and does not reflect recent events such as xAI's Grok 4 release, Meta's superintelligence announcement, or OpenAI's commitment to sign the EU AI Act Code of Practice.*

These findings reveal an unregulated industry where competitive pressures and technological ambition far outpace safety infrastructure and norms. This imbalance becomes more dangerous as companies pursue their stated goal of achieving artificial general intelligence (AGI) that matches or exceeds human capabilities within the decade.

4.2 Improvement opportunities by company

Here we highlight examples of how individual companies can improve future scores with relatively modest effort.

Anthropic:

- Publish a full whistleblowing policy to match OpenAI's transparency standard.
- Become more transparent and explicit about risk assessment methodology—e.g. why/how exactly is the particular eval related to a (class of) risks. Include reasoning in model cards that explicitly links evaluations or experimental procedures to specific risk, with limitations and qualifications.

OpenAI:

- Rebuild lost safety team capacity and demonstrate renewed commitment to OpenAI's original mission.
- Maintain the strength of current non-profit governance elements to guard against financial pressures undermining OpenAI's mission.

Google DeepMind:

- Publish a full whistleblowing policy to match OpenAI's transparency standard.
- Publish evaluation results for models without safety guardrails to more closely approximate true model capabilities.
- Improve coordination between DeepMind safety team and Google's policy team.
- Increase transparency around and investment in third-party model evaluations for dangerous capabilities.

xAI:

- Ramp up risk assessment efforts and publish implemented evaluations in upcoming model cards.
- Boost current draft safety framework to match the efforts by Anthropic and OpenAI.
- Publish a full whistleblowing policy to match OpenAI's transparency standard.

Meta:

- Significantly increase investment in technical safety research, especially tamper-resistant safeguards for open-weight models.
- Ramp up risk assessment efforts and publish implemented evaluations in upcoming model cards.
- Publish a full whistleblowing policy to match OpenAI's transparency standard.

Zhipu AI:

- Publish the AI Safety Framework promised at the AI Summit in Seoul.
- Ramp up risk assessment efforts and publish implemented evaluations in upcoming model cards.

DeepSeek:

- Address extreme jailbreak vulnerability before next release.
- Ramp up risk assessment efforts and publish implemented evaluations in upcoming model cards.
- Develop and publish a comprehensive AI safety framework.

All companies: Publish a first concrete plan, however imperfect, for how they hope to control the AGI/ASI they plan to build.

4.3 Domain-level findings

✂ Risk Assessment

This domain evaluates the rigor and comprehensiveness of companies' risk identification and assessment processes for their current flagship models. The focus is on implemented assessments, not stated commitments.

	 Anthropic	 OpenAI	 Google DeepMind	 x.AI	 Meta	 Zhipu AI	 DeepSeek
Domain Grade	C+	C	C-	F	D	F	F
Score	2.5	2.2	1.8	0	1.0	0.35	0

Indicator overview

Internal Testing

Dangerous Capability Evaluations
Elicitation for Dangerous Capability Evaluations
Human Uplift Trials

External Testing

Independent Review of Safety Evaluations
Pre-deployment External Safety Testing
Bug Bounties for Model Vulnerabilities

Only three companies—Anthropic, Google DeepMind, and OpenAI—were found to show meaningful efforts to assess whether their models pose large-scale risks. Reviewers recognized these efforts as demonstrating a substantial investment, highlighting Anthropic's and OpenAI's assessment of helpful-only models without safety guardrails as notable best practice. The review panel furthermore commended Anthropic as the only company to conduct a human participant uplift trial to evaluate the impact of its flagship model on risks from bioterrorism. The review panel found that the remaining four companies lack basic risk assessment documentation for critical risks, with basic practices such as dangerous capability evaluations and model cards mostly absent.

However, even the leaders were **not deemed to be sufficiently rigorous** in their approaches by the review panel. One expert criticized the lack of methodological transparency, noting: "The methodology/reasoning explicitly linking a given evaluation or experimental procedure to the risk, with limitations and qualifications, is usually absent." The reviewer encouraged the companies to draw from the risk assessment literature and improve their approach by being more "transparent and explicit about their risk assessment methodology (e.g., why/how exactly is the particular eval related to a (class of) risks". Currently, reported assessments were found to feature little explanation of why specific evaluations were chosen, what risks they target, or how results should be interpreted.

Reviewers also expressed concerns that **none of the companies commissioned independent verifications or assessments of internal safety evaluations, which means reported evidence needs to be accepted on trust.**

While OpenAI and Anthropic tasked competent external evaluators to assess some risks, reviewers noted that these efforts provide limited assurance, as evaluators are made to sign NDAs. Highlighting the severity of industry-wide deficiencies, one panellist who, despite assigning Anthropic the highest score among all firms, concluded her assessment by stating: "I have very low confidence that dangerous capabilities are being detected in time to prevent significant harm. Minimal overall investment in external 3rd party evals decreases my confidence further."

Current Harms

Companies were evaluated on how effectively their models mitigate current harms, with a focus on safety benchmark performance, robustness against adversarial attacks, watermarking of AI-generated content, and the treatment of user data.

	Anthropic	OpenAI	Google DeepMind	x.AI	Meta	Zhipu AI	DeepSeek
Domain Grade	B-	B	C+	D+	D+	D	D-
Score	2.8	3.0	2.5	1.5	1.65	1.0	0.85

Indicator overview

Model Safety / Trustworthiness

Stanford's HELM Safety Benchmark
Stanford's HELM AIR Benchmark
TrustLLM Benchmark

Robustness

Gray Swan Arena: UK AISI Agent Red-Teaming Challenge
Cisco Security Risk Evaluation
Protecting Safeguards from Fine-tuning

Digital Responsibility

Watermarking
User Privacy

The review panel found that performance on safety benchmarks varies drastically, from B's to D's. All models remain vulnerable to jailbreaks and misuse, with DeepSeek standing out for particularly high failure rates in adversarial testing. Reviewers positively noted the incremental improvements in model robustness showcased by Anthropic and OpenAI. **The panel criticized Meta, ZhipuAI, and DeepSeek for further amplifying risks by fully releasing their model weights, which enables malicious actors to remove safety protections through fine-tuning.**

Watermarking systems were found to remain underdeveloped for most companies, despite their importance in addressing the harms of synthetic content. Google DeepMind's SynthID stood out as the most advanced implementation.

On privacy matters, Anthropic was highlighted as the only firm not to train on user interaction data by default. Reviewers acknowledged that Meta, Zhipu, and DeepSeek offer users the ability to self-host their AIs by sharing model weights, which enables the highest level of privacy.

Safety Frameworks

This domain evaluates the companies' published safety frameworks for frontier AI development and deployment from a risk management perspective. The [comprehensive analysis](#) for the indicators in this domain was conducted by the non-profit research organisation [SaferAI](#).

	Anthropic	OpenAI	Google DeepMind	x.AI	Meta	Zhipu AI	DeepSeek
Domain Grade	C	C	D+	D+	D+	F	F
Score	2.15	2.0	1.5	1.5	1.5	0	0

Indicator overview

Risk Identification

Risk Analysis and Evaluation

Risk Treatment

Risk Governance

ZhipuAI and DeepSeek were assigned Fs for not having published a comparable Safety Framework in the first place, even though ZhipuAI had promised to do so at the international AI summit in Seoul. The remaining firms

published frameworks, with reviewers highlighting distinct strengths: Anthropic’s risk identification and mitigation approach, OpenAI’s commitment to publishing evaluation results, Google DeepMind’s alert thresholds, and Meta’s provisions for ongoing monitoring and threat modeling. However, reviewers identified one overarching criticism: the very limited scope of risks addressed by these frameworks.

The absence of external oversight mechanisms has emerged as a fundamental weakness of current frameworks, with OpenAI and Anthropic being noted for their early efforts to include external stakeholders. One reviewer explained that they emphasized risk treatment and governance in their weighting because: “[..] if no one has point and power [...], the quality of risk understanding is kind of moot.” The panel further pointed out that no framework sufficiently defined specifics around conditional pauses. While firms signed the Seoul commitments, none have spelled out concrete, externally verifiable trigger thresholds for pauses, nor reliable enforcement mechanisms.

Overall, panellists concluded that none of the companies could be trusted to prevent catastrophic risks with a high degree of confidence. Experts found that while the quality of these voluntary frameworks is slowly improving, they lack critical governance mechanisms that can ensure frameworks are implemented and enforced in high-stakes situations.

SaferAI’s complete evaluation of firms’ safety frameworks contains additional analysis from risk management professionals and can be found [here](#). Differences in scoring are due to the weightings applied by our review panel.

Existential Safety

This domain examines companies’ preparedness for managing extreme risks from future AI systems that could match or exceed human capabilities, including stated strategies and research for alignment and control.

	Anthropic	OpenAI	Google DeepMind	x.AI	Meta	Zhipu AI	DeepSeek
Domain Grade	D	F	D-	F	F	F	F
Score	1.0	0.67	0.77	0.23	0.33	0	0

Indicator overview

Existential Safety Strategy

Internal Monitoring and Control Interventions

Technical AI Safety Research

Supporting External Safety Research

“This category is deeply disturbing,” one reviewer noted. All seven companies are racing to build AGI within the decade, yet “literally none of the companies has anything like a coherent, actionable plan for what should happen if what they say will happen soon and are very actively working to make happen, happens”. Multiple reviewers emphasized this stark disconnect between ambition and preparedness, with five of seven companies scoring an F, and none of them scoring better than a D.

Quantitative guarantees for alignment or control strategies were found to be virtually absent, with no firm providing formal safety proofs or probabilistic risk bounds for the transformative technologies they set out to develop. “Companies working on AGI need to show that risks are actually below an acceptable threshold. None of them have a plan to do this,” one reviewer highlighted, adding that even those showing awareness are pursuing approaches “unlikely to yield the necessary level of safety.”

Anthropic and Google DeepMind received a D and a D-. One reviewer highlighted Anthropic’s “world-leading research on scheming / alignment faking, which lends credibility to their commitment to detecting misalignment.” However, the panel criticised the firm’s strategy’s over-reliance on mechanistic interpretability, given that the discipline is in an early stage.








Google DeepMind’s safety documentation was described as “well thought out, with a serious commitment to monitoring,” but reviewers noted it provides no solid foundation for assessing risks as acceptably low.

OpenAI’s deteriorating safety culture drew particular concern, leading to a grade drop to an F. “OpenAI’s focus on safety has decreased over the last year, and it has lost most of its researchers in this area,” a reviewer noted. High turnover on the safety team and failure to meet Superalignment commitments were taken as an indication of a concerning shift in priorities.

While xAI’s and ZhipuAI’s leadership was acknowledged for showing awareness of catastrophic risks, the companies themselves produce little concrete technical research aimed at addressing these risks. DeepSeek was also found not to publish related technical research.

Governance and Accountability

This domain assesses whether each company’s governance structure and day-to-day operations prioritize meaningful accountability for the real-world impacts of its AI systems. Indicators examine whistleblowing systems, legal structures, and advocacy efforts related to AI regulations.

	 Anthropic	 OpenAI	 Google DeepMind	 x.AI	 Meta	 Zhipu AI	 DeepSeek
Domain Grade	A-	C-	D	C-	D-	D+	D+
Score	3.7	1.7	1.0	1.85	0.85	1.35	1.35

Indicator overview

Lobbying on AI Safety Regulations
Company Structure & Mandate

Whistleblowing
Whistleblowing Policy Transparency
Whistleblowing Policy Quality Analysis
Reporting Culture & Whistleblowing Track Record

Anthropic stood out to reviewers for its Public Benefit Corporation status and Long-Term Benefit Trust, designed to reduce short-term profit incentives and strengthen long-term safety considerations. The review panel assessed OpenAI’s planned transition away from its original non-profit structure as weakening alignment to its safety-focused mission. xAI is also registered as a Public Benefit Corporation, but has not yet demonstrated how that structure translates into meaningful safety governance.








The review panel weighted advocacy efforts related to AI regulations as a significant factor in their assessments. Panel members noted Anthropic’s partial endorsement of SB 1047 and its opposition to a federal preemption of state-level regulation. x.AI’s CEO was acknowledged for publicly supporting SB 1047. In contrast, experts reduced grades for OpenAI, Google DeepMind, and Meta for lobbying against key AI safety regulations, including the EU AI Act, SB 1047, and the RAISE Act.

The review panel identified robust public whistleblowing policies as an industry-wide gap, with panel members noting that OpenAI was the only company to publish its whistleblowing policy—a standard the panel considered a

basic expectation in safety-critical industries. While this transparency was seen as a positive step that other firms should emulate, OpenAI has also drawn criticism for its previous use of restrictive non-disclosure agreements tied to the vested equity of former employees. Experts flagged Google DeepMind and Meta for past incidents involving retaliation against employees who raised concerns, which panel members assessed as potentially having a chilling effect on safety culture.

Information Sharing

This section gauges how openly firms share information about products, risks, and risk management practices. Indicators cover voluntary cooperation, transparency on technical specifications, and risk/incident communication.

	 Anthropic	 OpenAI	 Google DeepMind	 x.AI	 Meta	 Zhipu AI	 DeepSeek
Domain Grade	A-	B	B	C+	D	D	F
Score	3.7	3.0	3.0	2.3	1.0	1.0	0
Indicator overview							
Technical Specifications		Voluntary Cooperation		Risks & Incidents			
System Prompt Transparency		G7 Hiroshima AI Process Reporting		Serious Incident Reporting & Government Notifications			
Behavior Specification Transparency		FLI AI Safety Index Survey Engagement		Extreme-Risk Transparency & Engagement			

Compared to last year, OpenAI stood out for its engagement with the AI Safety Index survey. The firm's detailed model specification and regular disclosure of identified instances of malicious misuse were positively highlighted by reviewers.

Risk communication was found to diverge sharply between firms. Reviewers recognized Anthropic's proactive stance in informing policymakers and the public about critical risks, while Meta lost scores because its leadership publicly downplays extreme risks.

System prompt secrecy was found to still be the norm for proprietary models, with only Anthropic and xAI receiving credit for exposing the texts that steer their models.

Reviewers noted that incident reporting frameworks are largely absent, with none of the seven companies providing a concrete, public process for notifying governments about critical incidents. It was noted that this absence could undermine collective learning, and slow government responses to emerging threats.

xAI, ZhipuAI, and OpenAI received credit for submitting responses to the AI Safety Index survey. Regarding voluntary cooperation with the G7 Hiroshima AI Reporting Process, Meta and xAI stood out to reviewers as the only firms based in G7 countries that did not submit documentation on their risk management system.

5 Conclusions

The 2025 FLI AI Safety Index reveals an industry trust crisis: despite growing international consensus on AI risks and mounting evidence of rapid capability advances, experts warn that the gap between technological ambition and safety preparedness is widening. With no company achieving a grade higher than a C+ overall, reviewers expressed doubts that the industry's self-regulatory approaches will prove sufficient to address the magnitude of the risks.

The report's findings paint a troubling picture: Companies are racing toward artificial general intelligence and predict they will achieve superhuman performance within this decade. Yet as one reviewer noted, *"none of the companies has anything like a coherent, actionable plan"* for controlling such systems. While some firms invested in meaningful evaluations of dangerous capabilities despite fierce competitive pressure, reviewers still identified glaring methodological shortcomings across the industry. One expert warned,

"I have very low confidence that dangerous capabilities are being detected in time to prevent significant harm. Minimal overall investment in external 3rd party evaluations decreases my confidence further."

The Future of Life Institute remains committed to tracking these critical developments through regular Index updates. We will continue working with our expert review panel and partner organizations to refine our assessments and highlight both concerning gaps and emerging best practices.