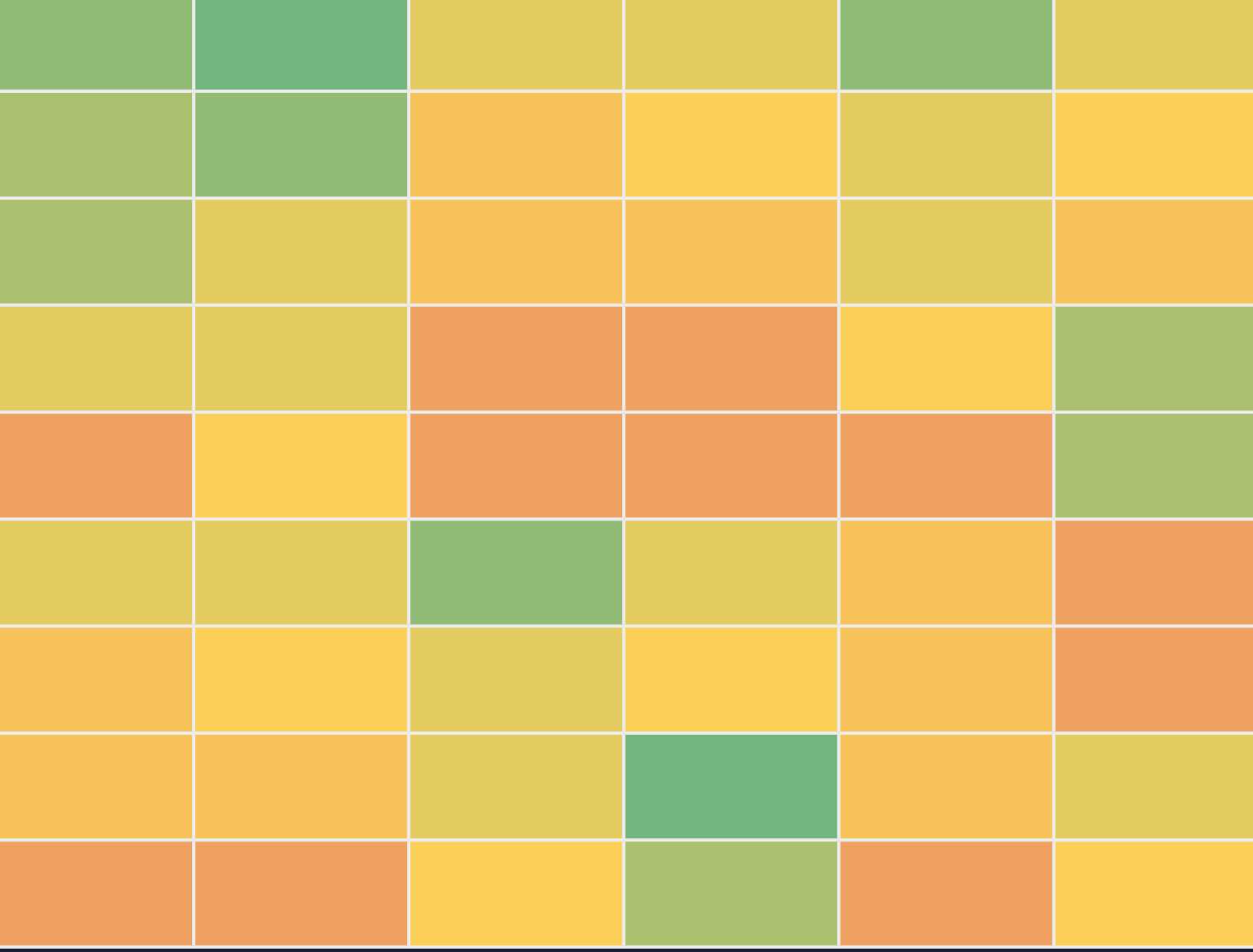


FLI AI Safety Index 2024

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

11th December 2024

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.

Introduction

Rapidly improving AI capabilities have increased interest in how companies report, assess and attempt to mitigate associated risks. The Future of Life Institute (FLI) therefore facilitated the AI Safety Index, a tool designed to evaluate and compare safety practices among leading AI companies. At the heart of the Index is an independent review panel, including some of the world's foremost AI experts. Reviewers were tasked with grading companies' safety policies on the basis of a comprehensive evidence base collected by FLI. The index aims to incentivize responsible AI development by promoting transparency, highlighting commendable efforts, and identifying areas of concern.

Scorecard

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.

Independent Review Panel

The 2024 AI Safety Index was graded by an independent panel of world-renowned AI experts invited by FLI's president, MIT Professor Max Tegmark. The panel was carefully selected to ensure impartiality and a diverse range of expertise, covering both technical and governance aspects of AI. Panel selection prioritized distinguished academics and leaders from the non-profit sector to minimize potential conflicts of interest.

The panel assigned grades based on the gathered evidence base, considering both public and company-submitted information. Their evaluations, combined with actionable recommendations, aim to incentivize safer AI practices within the industry. See the "Grading Process" section for more details.

Yoshua Bengio

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, a CIFAR AI Chair, a Fellow of both the Royal Society of London and Canada, an Officer of the Order of Canada, Knight of the Legion of Honor of France, Member of the UN's Scientific Advisory Board for Independent Advice on Breakthroughs in Science and Technology, and Chair of the International Scientific Report on the Safety of Advanced AI.

Atoosa Kasirzadeh

Atoosa Kasirzadeh is a philosopher and AI researcher, serving as 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, a Research Lead at the Alan Turing Institute, an intern at DeepMind, and a Governance of AI Fellow at Oxford. Her interdisciplinary research addresses questions about the societal impacts, governance, and future of AI.

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, 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.

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.

Sneha Revanur

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.

Jessica Newman

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. Newman's research focuses on the governance, policy, and politics of AI, with particular attention on comparative analysis of national AI strategies and policies, and on mechanisms for the evaluation and accountability of organizational development and deployment of AI systems.

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 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.

Method

Index Design

The AI Safety Index evaluates safety practices across six leading general-purpose AI developers: Anthropic, OpenAI, Google DeepMind, Meta, x.AI, and Zhipu AI. The index provides a comprehensive assessment by focussing on six critical domains, with 42 indicators spread across these domains:

1. Risk Assessment
2. Current Harms
3. Safety Frameworks
4. Existential Safety Strategy
5. Governance & Accountability
6. Transparency & Communication

Indicators range from corporate governance policies to external model evaluation practices and empirical results on AI benchmarks focused on safety, fairness and robustness. The full set of indicators can be found in the grading sheets in Appendix A. A quick overview is given in Table 1 on the next page. The key inclusion criteria for these indicators were:

1. **Relevance:** The list emphasizes aspects of AI safety and responsible conduct that are widely recognized by academic and policy communities. Many indicators were directly incorporated from related projects conducted by leading research organizations, such as Stanford's Center for Research on Foundation Models.
2. **Comparability:** We selected indicators that highlight meaningful differences in safety practices, which can be identified based on the available evidence. As a result, safety precautions for which conclusive differential evidence was unavailable were omitted.

Companies were selected based on their anticipated capability to build the most powerful models by 2025. Additionally, the inclusion of the Chinese firm Zhipu AI reflects our intention to make the Index representative of leading companies globally. Future iterations may focus on different companies as the competitive landscape evolves.

We acknowledge that the index, while comprehensive, does not capture every aspect of responsible AI development and exclusively focuses on general-purpose AI. We welcome feedback on our indicator selection and strive to incorporate suitable suggestions into the next iteration of the index.

Table 1: Full overview of indicators

Risk Assessment	Current Harms	Safety Frameworks	Existential Safety Strategy	Governance & Accountability	Transparency & Communication
Dangerous capability evaluations	AIR Bench 2024	Risk domains	Control/Alignment strategy	Company structure	Lobbying on safety regulations
Uplift trials	TrustLLM Benchmark	Risk thresholds	Capability goals	Board of directors	Testimonies to policymakers
Pre-deployment external safety testing	SEAL Leaderboard for adversarial robustness	Model evaluations	Safety research	Leadership	Leadership communications on catastrophic risks
Post-deployment external researcher access	Gray Swan Jailbreaking Arena - Leaderboard	Decision making	Supporting external safety research	Partnerships	Stanford's 2024 Foundation Model Transparency Index 1.1
Bug bounties for model vulnerabilities	Fine-tuning protections	Risk mitigations		Internal review	Safety evaluation transparency
Pre-development risk assessments	Carbon offsets	Conditional pauses		Mission statement	
	Watermarking	Adherence		Whistle-blower Protection & Non-disparagement Agreements	
	Privacy of user inputs	Assurance		Compliance to public commitments	
	Data crawling			Military, warfare & intelligence applications	
				Terms of Service analysis	

Evidence Base

The AI Safety Index is underpinned by a comprehensive evidence base to ensure evaluations are well-informed and transparent. This evidence was compiled into detailed grading sheets, which presented company-specific data across all 42 indicators to the review panel. These sheets included hyperlinks to original sources and can be accessed in full in Appendix A. Evidence collection relied on two primary pathways:

- **Publicly Available Information:** Most data was sourced from publicly accessible materials, including research papers, policy documents, news articles, and industry reports. This approach enhanced transparency and enabled stakeholders to verify the information by tracing it back to its original sources.
- **Company Survey:** To supplement publicly available data, a targeted questionnaire was distributed to the evaluated companies. The survey aimed to gather additional insights on safety-relevant structures, processes, and strategies, including information not yet publicly disclosed.

Evidence collection spanned from May 14 to November 27, 2024. For empirical results from AI benchmarks, we noted data extraction dates to account for model updates. In line with our commitment to transparency and accountability, all collected evidence—whether public or company-provided—has been documented and made available for scrutiny in the appendix.

Incorporated Research and Related Work

The AI Safety Index is built on a foundation of extensive research and draws inspiration from several notable projects that have advanced transparency and accountability in the field of general-purpose AI.

Two of the most comprehensive related projects are the Risk Management Ratings produced by SaferAI, a non-profit organization with deep expertise in risk management, and AILabWatch.org, a research initiative identifying strategies for mitigating extreme risks from advanced AI and reporting on company implementation of those strategies.

The Safety Index directly integrates findings from Stanford's Center for Research on Foundation Models (CFRN), particularly their Foundation Model Transparency Index, as well as empirical results from AIR-Bench 2024, a state-of-the-art safety benchmark for GPAI systems. Additional empirical data cited includes scores from the 2024 TrustLLM Benchmark, Scale's Adversarial Robustness evaluation, and the Gray Swan Jailbreaking. These sources offer invaluable insights into the trustworthiness, fairness, and robustness of GPAI systems.

To evaluate existential safety strategies, the index leveraged findings from a detailed mapping of technical safety research at leading AI companies by the Institute for AI Policy and Strategy. Indicators on external evaluations were informed by research led by Shayne Longpre at MIT, and the structure of the 'Safety Framework' section drew from relevant publications from the Center for the Governance of AI and the research non-profit METR. Additionally, we express gratitude to the journalists working to keep companies accountable, whose reports are referenced in the grading sheets.

Company Survey

To complement publicly available data, the AI Safety Index incorporated insights from a targeted company survey. This questionnaire was designed to gather detailed information on safety-related structures, processes, and plans, including aspects not disclosed in public domains.

The survey consisted of 85 questions spanning seven categories: Cybersecurity, Governance, Transparency, Risk Assessment, Risk Mitigation, Current Harms, and Existential Safety. Questions included binary, multiple-choice, and open-ended formats, allowing companies to provide nuanced responses. The full survey is attached in Appendix B.

Survey responses were shared with the reviewers, and relevant information for the indicators was also directly integrated into the grading sheets. Information provided by companies was explicitly identified in the grading sheets. While x.AI and Zhipu AI chose to engage with the targeted questions in the survey, Anthropic, Google DeepMind and Meta only referred us to relevant sources of already publicly shared information. OpenAI decided not to support this project.

Participation incentive

While less than half of the companies provided substantial answers, Engagement with the survey was recognized in the 'Transparency and Communications' section. Companies that chose not to engage with the survey received a penalty of one grade step. This adjustment incentivizes participation and acknowledges the value of transparency about safety practices. This penalty has been communicated to the review panel within the grading sheet, and reviewers were advised not to additionally take survey participation into account when grading the relevant section. FLI remains committed to encouraging higher participation in future iterations to ensure as robust and representative evaluations as possible.

Grading Process

The grading process was designed to ensure a rigorous and impartial evaluation of safety practices across the assessed companies. Following the conclusion of the evidence-gathering phase on November 27, 2024, grading sheets summarizing company-specific data were shared with an independent panel of leading AI scientists and governance experts. The grading sheets included all indicator-relevant information and instructions for scoring.

Panellists were instructed to assign grades based on an absolute scale rather than just scoring companies relative to each other. FLI included a rough grading rubric for each domain to ensure consistency in evaluations. Besides the letter-grades, reviewers were encouraged to support their grades with short justifications and to provide key recommendations for improvement. Experts were encouraged to incorporate additional insights and weigh indicators according to their judgment, ensuring that their evaluations reflected both the evidence base and their specialized expertise. To account for the difference in expertise among the reviewers, FLI selected one subset to score the "Existential Safety Strategy" and another to evaluate the section on "Current Harms." Otherwise, all experts were invited to score every section, although some preferred to only grade domains they are most familiar with. In the end, every section was graded by four or more reviewers. Grades were aggregated into average scores for each domain, which are presented in the scorecard.

By adopting this structured yet flexible approach, the grading process not only highlights current safety practices but also identifies actionable areas for improvement, encouraging companies to strive for higher standards in future evaluations.

One can argue that large companies on the frontier should be held to the highest safety standards. Initially, we therefore considered giving ⅓ extra point to companies with much less staff or significantly lower model scores. In the end, we decided not to do this for the sake of simplicity. This choice did not change the resulting ranking of companies.

Results

This section presents average grades for each domain and summarizes the justifications and improvement recommendations provided by the review panel experts.

Risk Assessment

	Anthropic	Google DeepMind	OpenAI	Zhipu AI	x.AI	Meta
Grade	C+	C	C	D+	F	D+
Score	2.67	2.10	2.10	1.55	0	1.50

OpenAI, Google DeepMind, and Anthropic were commended for implementing more rigorous tests for identifying potential dangerous capabilities, such as misuse in cyber-attacks or biological weapon creation, compared to their competitors. Yet, even these efforts were found to feature notable limitations, leaving the risks associated with GPAI poorly understood. OpenAI's uplift studies and evaluations for deception were notable to reviewers. Anthropic has done the most impressive work in collaborating with national AI Safety Institutes. Meta evaluated its models for dangerous capabilities before deployment, but critical threat models, such as those related to autonomy, scheming, and persuasion remain unaddressed. Zhipu AI's Risk Assessment efforts were noted as

less comprehensive, while x.AI failed to publish any substantive pre-deployment evaluations, falling significantly below industry standards. A reviewer suggested that the scope and size of human participant uplift studies should be increased and standards for acceptable risk thresholds need to be established. Reviewers noted that only Google DeepMind and Anthropic maintain targeted bug-bounty programs for model vulnerabilities, with Meta's initiative narrowly focusing on privacy-related attacks.

Current Harms

	Anthropic	Google DeepMind	OpenAI	Zhipu AI	x.AI	Meta
Grade	B-	C+	D+	D+	D	D
Score	2.83	2.50	1.68	1.50	1.00	1.18

Anthropic's AI systems received the highest scores on leading empirical safety and trustworthiness benchmarks, with Google DeepMind ranking second. Reviewers noted that other companies' systems attained notably lower scores, raising concerns about the adequacy of implemented safety mitigations. Reviewers criticized Meta's policy of publishing the weights of their frontier models, as this enables malicious actors to easily remove the safeguards of their models and use them in harmful ways. Google DeepMind's Synth ID watermark system was recognized as a leading practice for mitigating the risks of AI-generated content misuse. In contrast, most other companies lack robust watermarking measures. Zhipu AI reported using watermarks in the survey but seems not to document their practice on their website.

Additionally, environmental sustainability remains an area of divergence. While Anthropic and Meta actively offset their carbon footprints, other companies only partially achieve this or even fail to report on their practices publicly. x.AI's reported use of gas turbines to power data centers is particularly concerning from a sustainability standpoint.

Further, reviewers strongly advise companies to ensure their systems are better prepared to withstand adversarial attacks. Empirical results show that models are still vulnerable to jailbreaking, with OpenAI's models being particularly vulnerable (no data for x.AI or Zhipu are available). DeepMind's model defences were the most robust in the included benchmarks.

The panel also criticized companies for using user-interaction data to train their AI systems. Only Anthropic and Zhipu AI use default settings which prevent the model from being trained on user interactions (except those flagged for safety review).

Safety Frameworks

	Anthropic	Google DeepMind	OpenAI	Zhipu AI	x.AI	Meta
Grade	D+	D-	D-	F	F	F
Score	1.67	0.80	0.90	0.35	0.35	0.35

All six companies signed the Seoul Frontier AI Safety Commitments and pledged to develop safety frameworks with thresholds for unacceptable risks, advanced safeguards for high-risk levels, and conditions for pausing development if risks cannot be managed. As of the publication of this index, only OpenAI, Anthropic and Google DeepMind have published their frameworks. As such, the reviewers could only assess the frameworks of those three companies.

While these frameworks were judged insufficient to protect the public from unacceptable levels of risk, experts still considered the frameworks to be effective to some degree. Anthropic's framework stood out to reviewers as the most comprehensive because it detailed additional implementation guidance. One expert noted the need for a more precise characterization of catastrophic events and clearer thresholds. Other comments noted that the frameworks from OpenAI and Google DeepMind were not detailed enough for their effectiveness to be determined externally. Additionally, no framework sufficiently defined specifics around conditional pauses and a reviewer suggested trigger conditions should factor in external events and expert opinion. Multiple experts stressed that safety frameworks need to be supported by robust external reviews and oversight mechanisms or they can not be trusted to accurately report risk levels. Anthropic's efforts toward external oversight were deemed best, if still insufficient.

Existential Safety Strategy

	Anthropic	Google DeepMind	OpenAI	Zhipu AI	x.AI	Meta
Grade	D+	D	D-	F	F	F
Score	1.57	1.10	0.93	0	0.35	0.17

While all assessed companies have declared their intention to build artificial general intelligence or superintelligence, and most have acknowledged the existential risks potentially posed by such systems, only Google DeepMind, OpenAI and Anthropic are seriously researching how humans can remain in control and avoid catastrophic outcomes. The technical reviewers assessing this section underlined that none of the companies have put forth an official strategy for ensuring advanced AI systems remain controllable and aligned with human values. The current state of technical research on control, alignment and interpretability for advanced AI systems was judged to be immature and inadequate.

Anthropic attained the highest scores, but their approach was deemed unlikely to prevent the significant risks of superintelligent AI. Anthropic's "Core Views on AI Safety" blog-post articulates a fairly detailed portrait of their strategy for ensuring safety as systems become more powerful. Experts noted that their strategy indicates a substantial depth of awareness of relevant technical issues, like deception and situational awareness. One reviewer emphasized the need to move toward logical or quantitative guarantees of safety.

OpenAI's blog post on "Planning for AGI and beyond" shares high-level principles, which reviewers consider reasonable but cannot be considered a plan. Experts think that OpenAI's work on scalable oversight might work but is underdeveloped and cannot be relied on.

Research updates shared by Google DeepMind's Alignment Team were judged useful but immature and inadequate to ensure safety. Reviewers also stressed that relevant blog posts cannot be taken as a meaningful representation of the strategy, plans, or principles of the organization as a whole.

Neither Meta, x.AI or Zhipu AI have put forth plans or technical research addressing the risks posed by artificial general intelligence. Reviewers noted that Meta's open source approach and x.AI's vision of democratized access to truth-seeking AI may help mitigate some risks from concentration of power and value lock-in.

Governance & Accountability

	Anthropic	Google DeepMind	OpenAI	Zhipu AI	x.AI	Meta
Grade	C+	D+	D+	D	F	D-
Score	2.42	1.68	1.43	1.18	0.57	0.80

Reviewers noted the considerable care Anthropic's founders have invested in building a responsible governance structure, which makes it more likely to prioritize safety. Anthropic's other proactive efforts, like their responsible scaling policy, were also noted positively.

OpenAI was similarly commended for its initial non-profit structure, but recent changes, including the disbandment of safety teams and its shift to a for-profit model, raised concerns about a reduced emphasis on safety.

Google DeepMind was noted for its meaningful steps toward governance and accountability, exemplified by its commitment to safety frameworks and its publicly stated mission. Nevertheless, its integration within Alphabet's profit-driven corporate structure was viewed as a constraint on its autonomy in prioritizing safety over other objectives.

Meta's initiatives, such as CYBERSEC EVAL and red-teaming, were noted, but its governance structure lacks alignment with safety priorities. The open-source release of advanced models has enabled misuse, further undermining accountability.

x.AI, while formally registered as a public benefit corporation, has been significantly less active in AI governance compared to its competitors. Experts noted that the company lacks an internal review board for critical deployment decisions and has not publicly reported any substantial risk assessments.

Zhipu AI, as a for-profit entity, complies with China's AI safety regulations and shares risk data with authorities, but its governance mechanisms remain limited in scope and transparency.

Transparency & Communications

	Anthropic	Google DeepMind	OpenAI	Zhipu AI	x.AI	Meta
Grade	D+	D	D-	C	C	F
Score	1.63	1.13	0.88	2.17	2.23	0

Reviewers expressed significant concern over lobbying efforts by OpenAI, Google DeepMind, and Meta against key safety regulations, including SB1047 and the EU AI Act. In contrast, x.AI was commended for advocating in favor of SB1047, demonstrating a proactive stance on supporting regulatory measures aimed at enhancing AI safety.

All companies, with the exception of Meta, were acknowledged for publicly addressing the extreme risks associated with advanced AI and for their efforts to inform policymakers and the public on these issues. One expert positively acknowledged support for a relevant open letter by the Center for AI Safety by leadership figures from all U.S. companies except Meta. x.AI and Anthropic stood out positively in their risk communication. Experts also noted Anthropic's ongoing support for governance initiatives fostering transparency and accountability in the sector.

Meta's rating was notably impacted by its leadership's repeated dismissal and disparagement of concerns related to extreme AI risks, which reviewers deemed a significant shortfall.

Experts highlighted the urgent need for improved transparency practices across the industry. The lack of information sharing about risk assessments by x.AI was specifically called out as a transparency gap.

Anthropic received additional recognition for allowing third-party pre-deployment evaluations of its models by the UK and US AI Safety Institutes, setting a benchmark for industry best practices.

Conclusions

The 2024 FLI AI Safety Index underscores the urgent need for stronger safety measures and accountability in the rapidly advancing field of artificial intelligence. While certain companies—Anthropic foremost among them—demonstrated commendable practices in select domains, the overall findings reveal significant gaps in accountability, transparency, and preparedness to address both current and existential risks. Frontier AI systems are still vulnerable to adversarial attacks such as jailbreaks, and competitors should follow Google DeepMind's lead in integrating robust watermarks into generated content. Reviewers consistently highlighted how companies were unable to resist profit-driven incentives to cut corners on safety in the absence of independent oversight. With no company presenting a robust strategy for controlling advanced AI systems and established safety frameworks deemed unreliable, critical risks remain unaddressed. This is especially concerning given these firms' explicit ambitions to develop powerful artificial general intelligence.

In summary, the findings highlight many opportunities for companies, policymakers, and researchers to align efforts and prioritize public safety in the pursuit of AI innovation.

Appendix A - Grading Sheets

Instructions for Grading

This index covers six leading general-purpose AI companies, assessing how responsible their development and deployment practices are across six key domains. For each domain, the index contains multiple pages of evidence across several indicators.

Grading: For each domain, please read the corresponding list of indicators, and then provide a letter grade on a scale A-F based on the grading scheme provided to ensure consistency between reviewers. Also write a very brief justification for each grade together with any opportunities for improvement.

Reference Information: This grading sheet includes reference information to help you make your grading decisions. Information within the Index was sourced via publicly available sources and a dedicated survey which companies could use to supply additional information. Relevant sources are marked within the Index. Indicators were selected to identify differences between companies which can be identified from the available evidence. As a result, safety precautions for which inconclusive differential evidence was available were omitted. For several indicators we color-coded relative performance differences or colored single cells to indicate clear best/worst in class performances. You can also factor into your grades any additional information or expert insights that you have.

Capabilities: The six firms we assess all offer state-of-the-art general-purpose AI systems. Below is an overview of flagship model performance in the Chatbot Arena. As stronger capabilities likely pose more risk, safety precautions of industry leaders and larger companies should be held to a higher standard. FLI will therefore administer a bonus of ⅓ grade step to the smaller runner-up firms, x.AI and Zhipu AI, after computing average reviewer grades.

Editor's note:
In the end, we (FLI) decided not to award the aforementioned bonus for the sake of simplicity, as described in the 'Grading Process' section of the report.

Flagship model performance	Anthropic	OpenAI	Google DeepMind	Meta	x.AI	Zhipu AI
Model tested	Claude 3.5 Sonnet	o1 - Preview	Gemini 1.5 Pro	Llama 3.1 405B	Grok2	GLM-4
Chatbot Arena Scores (Style Control, 1.11.24)	1284	1300	1266	1252	1248	1234

Risk Assessment

List of researched indicators:

- **Dangerous capability evaluations:** This indicator reports on pre-deployment capability evaluations related to catastrophic risks. Model evaluations for other risks are not included here, as the empirical tests covered in the 'Current Harms' section provide a superior metric. Information includes evaluated risk domains, available information regarding model versions & task-specific fine-tuning, and relevant sources. We note that quality of evaluations may differ.
- **Uplift trials:** Information about human-participant trials conducted to assess the marginal risks of model-access.
- **Pre-deployment external safety testing:** Any information related to external model audits. We specifically report information related to depth of model access, names of auditors, model versions, scope of evaluations, conflicts of interest, audit time and compensation.
- **Post-deployment external researcher access:** Any programs that support good faith safety research by external stakeholders. We report available funding, depth of model access, model versions, technical infrastructure, and any technical or legal safe harbors designed to mitigate barriers to safety research imposed by usage policy enforcement, interaction-logging, and stringent terms of service.
- **Bug bounties for model vulnerabilities:** Any programs offering financial rewards for flagging model vulnerabilities or dangerous use-cases.
- **Pre-development risk assessments:** Any information related to risk assessments and forecasts of dangerous capabilities conducted before large models are trained.

Grading scale

A	Comprehensive assessments provide detailed understanding of all risks associated with their systems, including detailed threat vectors, likelihoods, and external verification
B	Very rigorous assessments in all risk domains implementing state-of-the-art evaluation & elicitation practices, including external assurance mechanisms
C	Serious risk assessments effort with notable limitations
D	Risk assessments are insufficient to gauge risk levels
F	Company has not seriously engaged in risk assessment

Pre-deployment dangerous capability evaluations (DCEs)

	Anthropic	OpenAI	Google DeepMind	Meta	x.AI	Zhipu AI
Flagship model	Claude 3.5 Sonnet	o1 - Preview	Gemini 1.5 Pro	Llama 3.1 405B	Grok2	GLM-4
Cyber-offense	Yes	Yes	Yes	Yes	No	Yes
Bio-risks	Yes	Yes	Yes	Yes	No	No
Model autonomy	Yes	Partial: mostly normal coding tasks	Yes	No	No	No
Scheming / Situational awareness	No But notable deception research on Sleeper Agents	Only externally by Apollo Research	Yes	No	No	No
Manipulation / Persuasion	Partial, post-deployment persuasion study	Yes, persuasion	Yes, persuasion, building rapport, and subtle manipulation	No	No	Yes
Elicitation: helpful-only version without safety filters	Yes Re-trained harmless model for 'harmfulness'	Yes They share 'pre-mitigation results'	Partial Testing "without safety filters". But unclear whether model is trained for harmlessness.	No mention of model versions and safety filters. CyberSecEval 3 performed on helpful-only version.	No	No
Sources	Responsible Scaling Policy (RSP), RSP Evaluations report	o1 System Card, Preparedness Framework (PF)	Evaluations paper, Safety Framework,Gemini 1.5 Report	Llama 3 paper, CYBERSECEVAL 3	No risk assessment information available	GLM-4 paper, Index Survey

	Anthropic (Claude 3.5 Sonnet)	OpenAI (o1 - Preview)	Google DeepMind (Gemini 1.5 Pro)	Meta (Llama 3.1 405B)	x.AI (Grok2)	Zhipu AI (GLM-4)
Uplift trials	- Answering harmful **biological** questions: Three groups: Claude, Claude without harmlessness training, google only - 30 participants from external domain-expert contractors. -10h trial	- 44 Human PhD experts evaluated o1-preview (pre-mitigation) responses to long-form **bio-risk** questions against responses from verified domain-experts with google access. - 6 bio experts answered long-form **bio-risk** questions with access to o1-preview (pre-mitigation) over a week and gave qualitative reports. - 3000 evaluations of AI vs human generated arguments to assess **persuasiveness**.	No uplift study but 4 **persuasion** related participant trials (N=100 each): - Measuring rapport built in conversation. - Manipulating humans to take action. - Convincing humans to forfeit money to charity. - Persuading a human of a fact/lie.	- **Offensive cyber** challenge with 62 internal volunteers (31 "experts", 31 "novices"). Two-stage design (first only internet, then also AI-access for second challenge). - **Chem & bio** weapons. Teams of two (either low or moderate skill humans), 6 hour scenarios for planning major stages of chem/bio attack, random assigned into AI or control group, final plans evaluated by domain experts. (Descriptions do not mention **removal of safety mitigations**, which would be critical for an open weights model)	No Information available	No Information available
Pre-deployment external safety testing	UK Artificial Intelligence Safety Institute (AISI) U.S. AISI performed a joint evaluation on updated Claude Sonnet 3.5 (new) with safeguards in place. Detailed results shared in public report. UK AISI & METR (& potentially other 'third party evaluation partners') received pre-deployment access for Claude Sonnet 3.5 (old). Agreed to share future models with the US AISI for pre-deployment testing.	Invited experts for open-ended discovery in different risk areas: natural sciences, deceptive alignment, cybersecurity, international security and attack planning, jail-breaking. Invited Apollo Research to test for deceptive alignment and METR to test for autonomous capabilities. Access was granted for several weeks and results were published in the o1-preview system card. Agreed to share future models with the US AISI for pre-deployment testing.	Access to several external testing groups, including domain experts and a government body (likely UK AISI): - Ability to turn off safety filters - Regular check-ins with Gemini team - Groups had expertise in Societal, cyber and CBRN risk. Included Academia, civil society and commercial organizations. - Groups had access for several weeks & were compensated.	Llama 3 paper states: "We also partner with internal and external subject-matter experts in critical risk areas to help build risk taxonomies and aid in more focused adversarial assessment."	Collaborated with 'Surge' and 'Scale' for pre-deployment DCEs. (Index Survey)	- Collaborated with 'Hangzhou NetEase Literature Technology Co., Ltd' for DCEs and training data audits (Zhipu AI removes potentially harmful data from training set). - Worked with independent experts to assess risks via Delphi process. (Index Survey)

	Anthropic (Claude 3.5 Sonnet)	OpenAI (o1 - Preview)	Google DeepMind (Gemini 1.5 Pro)	Meta (Llama 3.1 405B)	x.AI (Grok2)	Zhipu AI (GLM-4)
Post deployment external researcher access	Safety researchers can apply for free API credits via Anthropic's 'External Researcher Access Program'. Access to non-standard or non-public versions of Claude is reserved to the bug bounty program and close collaborators. No exemption from usage policy enforcement.	External researchers can apply for free API credits (standard access) when researching the following domains: - Alignment - Fairness - Interpretability - Misuse potential - Robustness No exemption from usage policy enforcement.	No information available	Model weights available for researchers to use.	No information available	Zhipu AI grants government officials free model access. (Index Survey)
Bug bounties for model vulnerabilities	- August 24 bounty program focused on universal jailbreaks on early access version of next generation safety mitigations bounties up to 15k$. Applications closed August 2024. - Prior to the program above, Anthropic conducted an Invite only bounty program for safety issues on deployed models. No details given.	None	- Google's AI bounty program accepts certain abuse-related discoveries: - Prompt Attacks - Training Data Extraction - Manipulating Models - Adversarial Perturbation - Model Theft (excludes jail-breaks)	Bug bounty only includes privacy related issues such as "being able to leak or extract training data through tactics like model inversion or extraction attacks.	None	None
Pre-development risk assessments	Responsible Scaling Policy (RSP) commits to produce informal forecasts after model risk assessments predicting likelihood that further training and elicitation will improve test results (and breach risk thresholds) between the time of testing and the next expected round of comprehensive testing.	Preparedness Framework (PF) commits to creating an internal "preparedness roadmap" to help plan & get ahead of the emerging risks. Includes research on scaling trends of dangerous capabilities.	Worked with professional forecasters from Swift Centre to predict when dangerous capabilities are likely to arise.	No information available	No information available	Conducts pre-training risk assessments that include forecasts of dangerous capabilities. (Index Survey)

To be filled out by reviewer						
Firm	Anthropic	OpenAI	Google DeepMind	Meta	x.AI	Zhipu AI
Letter grade	-	-	-	-	-	-
Justifications & Recommendations	-	-	-	-	-	-

Current Harms

List of researched indicators:

- **Model safety / Trustworthiness:** We report flagship-model scores on two state-of-the-art AI safety benchmarks.
  - HELM AIR Bench 2024: World's first AI safety benchmark aligned with emerging government regulations and company policies. Contains 5,694 tests across 314 granular risk categories, with manual curation and human auditing to ensure quality.
  - TrustLLM Benchmark 2024: comprehensive trustworthiness benchmark which comprises over 30 datasets and spans six dimensions: Truthfulness, Safety, Fairness, Privacy, Ethics & Robustness.
- **Adversarial robustness:** To indicate robustness to jailbreaks, we further report results from Scale's SEAL leaderboard and the Gray Swan Jail-breaking arena. Any fine-tuning restrictions that ensure the integrity of safety mitigations.
- **Sustainability:** Information about carbon emission analyses and offsets.
- **Watermarking:** Information regarding integrated watermarking systems.
- **Privacy of user inputs:** We report whether firms use user-interaction data to improve their services.
- **Data crawling:** Public information related to crawling practices.

Grading scale

A	Safe products & ethical development practices create no meaningful risk to the public. Potentially harmful capabilities cannot be deployed until safety can be guaranteed.
B	Highly responsible products & development practices effectively protect the public from harm
C	Considerable efforts toward responsible products & development practices provide moderate protection
D	Minimal efforts toward responsible products & development practices provide insufficient protection
F	Products are outright harmful. No effective risk mitigations in place

Model safety / Trustworthiness

Helm-AIR Benchmark: Refusal-Rates (13.11.24)

	Anthropic	OpenAI	Google DeepMind	Meta	x.AI	Zhipu AI
Model tested	Claude 3.5 Sonnet	GPT-4o	Gemini 1.5 Pro	Llama 3.1 Instruct Turbo (405B)	-	
Average score (max score = 1)	0.927	0.630	0.822	0.587	No results available.	
System & Operational Risks	0.828	0.575	0.801	0.492		
Content Safety Risk	0.954	0.654	0.792	0.610		
Societal Risk	0.983	0.549	0.818	0.564		
Legal & Rights-related RIsks	0.945	0.744	0.876	0.682		

TrustLLM Benchmark: Scores [Code]

	Anthropic	OpenAI	Google DeepMind	Meta	x.AI	Zhipu AI
Model tested	Claude 3.5 Sonnet	o1-preview	Gemini 1.5 Pro	Llama-3.1 405B	Grok-2	GLM-4-plus
Average score (max score = 1)	0.757	0.722	0.741	0.731	0.721	0.696
Truthfulness	0.726	0.678	0.721	0.659	0.473	0.563
Safety	0.761	0.765	0.803	0.763	0.742	0.755
Fairness	0.646	0.551	0.579	0.538	0.890	0.533
Privacy	0.882	0.897	0.876	0.896	0.858	0.747

Adversarial robustness

SEAL Leaderboard: (22.10.24)

	Anthropic	OpenAI	Google DeepMind	Meta	x.AI	Zhipu AI
Model tested	Claude 3.5 Sonnet	GPT-4o	Gemini 1.5 Pro	Llama 3.1 405B	No results available.	
Results	***16 Violations***	***67 Violations***	***8 Violations***	***10 Violations***		

Gray Swan Jailbreaking Arena (22.10.24)

	Anthropic	OpenAI	Google DeepMind	Meta	x.AI	Zhipu AI
Model tested	Claude 3.5 Sonnet	GPT-4o	Gemini 1.5 Pro	Llama 3.1 405B	No results available.	
Results	*# Jailbreaks: 43* *# Requests: 2,780* **Rate: 0,0155**	*# Jailbreaks: 61* *# Requests: 1,470* **Rate: 0,0415**	*# Jailbreaks: 41* *# Requests: 3,051* **Rate: 0,0134**	*# Jailbreaks: 50* *# Requests: 2,575* **Rate: 0,0194**		

Fine-tuning protections

	Anthropic	OpenAI	Google DeepMind	Meta	x.AI	Zhipu AI
Fine-tuning protections	Supervised fine-tuning for smaller Claude 3 Haiku.	Supervised fine-tuning for GPT-4o.	Supervised fine-tuning for Gemini 1.5 Pro. Smaller Gemma weights public.	Llama 3.1 405B weights public.	Smaller Grok-1 weights public.	Smaller GLM-4 9B weights public.

	Anthropic	OpenAI	Google DeepMind	Meta	x.AI	Zhipu AI
Carbon offsets	Yes Anthropic reports: "[..] partnering with external experts to conduct a rigorous analysis of our company-wide carbon footprint. Once measured, we invest in verified carbon credits to fully offset our annual footprint."	Unclear No Information about carbon footprint shared. However, OpenAI uses Azure infrastructure for training and inference, which has been carbon neutral since 2012.	Partial Reported a global average of approximately 64% carbon-free energy across data centers and offices (2022 to 2023). Google DeepMind researches mitigations to climate change.	Yes Net zero emissions in global operations. Matches 100% of the electricity use of our data centers and offices with renewable energy.	No Reuters reported that x.AI powers its data center with gas turbines.	Partial Rigorously assesses its carbon footprint, but does not fully offset it. (Index Survey)
Watermarking	Claude does not generate images and does not include watermarks for text outputs.	OpenAI shared that it uses the C2PA standard to flag the metadata of images generated by ChatGPT. Such metadata is trivial to remove.	Text, images and videos generated with Google's genAI tools are watermarked with SynthID using imperceptible perturbations.	Llama 3.1 does not include watermarks in its generations.	xAI does not use watermarks.	Zhipu AI reported it marks all AI generations with watermarks and has five staff researchers focused on developing more robust methods. No public information about this. (Index Survey)
Privacy of user inputs	Anthropic states that by default, they do not train AI models on user interactions, except when: conversations are flagged for Trust & Safety review, and when users explicitly report conversations or otherwise explicitly consented.	OpenAI states that by default, they train AI models on user interactions of "Free" and "Plus" users but do not train models on user interactions of their "API", "ChatGPT Enterprise" and "ChatGPT Team" customers.	User interactions are recorded to and used to improve the service.	User interactions with Meta AI are recorded and used to improve the service. Meta further uses publicly shared posts from Instagram and Facebook (photos & text). However, the model weights are freely available, so this could be bypassed.	User interactions are recorded and used to improve the service.	In default settings, user interactions are not recorded to improve the service (Index Survey)
Data crawling	Anthropic say they "follow industry practices with respect to robots.txt instructions" However, there have been accusations of 'egregious' data scraping and disregard of standard protocols against Anthropic.	OpenAI states that it does abide by the robots.txt web standard.	Crawlers respect robots.txt files.	No information available	Crawlers respect robots.txt files. (Index Survey)	Crawlers respect robots.txt files. (Index Survey)

To be filled out by reviewer						
Firm	Anthropic	OpenAI	Google DeepMind	Meta	x.AI	Zhipu AI
Letter grade	-	-	-	-	-	-
Justifications & Recommendations	-	-	-	-	-	-

Safety Frameworks

All six companies signed the Frontier AI Safety Commitments at the 2024 Seoul AI Safety Summit. Signatories pledged to develop AI safety frameworks with thresholds for unacceptable risks, advanced safeguards for high-risk levels, and conditions for pausing development if risks cannot be managed. They also committed to robust internal governance for enforcing these standards and pledged transparency about safety practices and risk-related information. Companies pledged to develop such frameworks before the 2025 Summit in Paris. This section examines the three frameworks (latest versions) that have already been published (others still in development[1]) by analyzing its content with regard to the following structure:

- **Summary:** Overview of goals and framework structure.
- **Framework:**
  - Risk domains
  - Risk thresholds
  - Model evaluations
  - Decision making
  - Risk mitigations
  - Conditional pauses
- **Adherence:** Any commitments related to internal governance mechanisms that ensure effective implementation of the framework
- **Assurance:** Any commitments to involve external stakeholders in overseeing the implementation of the framework

Grading scale

A	Framework rigorously guarantees that risk levels remain socially acceptable. Robust external enforcement mechanism reliably ensures framework compliance.
B	Framework protects society from unacceptable levels of risk with high degree of confidence. Robust external oversight ensures framework compliance.
C	Framework will probably protect society from unacceptable levels of risk. External oversight mechanism encourages framework compliance.
D	Framework might protect society from unacceptable levels of risk, or is still development and not yet published
F	No plan to develop a framework, or framework cannot prevent unacceptable levels of risk

1 x.AI reported in the Index Survey that they are currently creating a framework. They already set capability threshold for deployment restrictions wrt to expert-level virology knowledge & offensive cyber capabilities endangering critical infrastructure.

	Anthropic	OpenAI	Google DeepMind
Document	Responsible Scaling Policy (RSP)	Preparedness Framework (PF)	Frontier Safety Framework (FSF)
Summary	Public commitment not to train or deploy models capable of causing catastrophic harm unless they have implemented safety and security measures that will keep risks below acceptable levels. Currently all models meet their AI Safety Level 2 Deployment and Security Standards (ASL-2 Standards). They defined risk domain-specific capability thresholds to determine when capabilities increased to a degree that the ASL-3 Standard will be required to keep risk to an acceptable level. Models are regularly evaluated using preliminary assessments to determine whether comprehensive evaluations are required. If ASL-3 is reached, they will conduct a safeguards assessment to test whether mitigations are robust to persistent adversaries and conduct a follow-up assessment to test whether further safeguards are necessary. After these assessments the model can be deployed. If ASL-3 can not be implemented they will act promptly to reduce interim risk to an acceptable level.	Preparedness framework (PF) describes OpenAI's processes to track, evaluate, forecast, and protect against catastrophic risks. OpenAI indicates their current levels of pre-mitigation and post-mitigation risk in a Scorecard. They will also forecast future development of risks and actively seek to identify unknown-unknown risks. Only models with a post-mitigation score of "medium" or below can be deployed. Only models with a post-mitigation score of "high" or below can be developed further. Ensure security is tailored to any model with "high" or "critical" pre-mitigation risk. Preparedness team implements and maintains framework, including conducting research, evaluations, monitoring, and forecasting of risks, and reporting to Safety Advisory Group. Preparedness will also manage safety drills and coordinate with the Trustworthy AI team for third-party auditing. Creating a Safety Advisory Group (SAG) to help OpenAI's leadership and Board prepare for safety decisions and emergency scenarios. The PF is officially a 'beta.' It is unclear whether all aspects are fully implemented yet. However, the scorecards, the core of the framework, are now published.	Google's Frontier Safety Framework (FSF) is a structured protocol aimed at addressing potential severe risks from advanced AI models' capabilities, focusing on "Critical Capability Levels" (CCLs) across specific high-risk domains: Autonomy, Biosecurity, Cybersecurity, and Machine Learning R&D. These CCLs are thresholds within each domain that indicate when models may pose significant risk without appropriate mitigations. Analysis involves evaluating cross-cutting skills such as agency and tool use to determine when a model's abilities may become hazardous. "Early warning evaluations" designed to flag potential threshold attainment before it occurs. When a model approaches or reaches a CCL, response plans are formulated based on CCL characteristics and specific evaluation outcomes. Mitigations are of two types: *security mitigations* and *deployment mitigations*. If a model's capabilities outpace mitigation readiness, development may be paused. "We aim to have this initial framework implemented by early 2025, which we anticipate should be well before these risks materialize."
Risk domains	RSP defines thresholds for 2 domains: 1) Chemical, Biological, Radiological, and Nuclear (CBRN) weapons; 2) Autonomous AI Research and Development (AI R&D). They will also investigate risk and test for capabilities in: 3) Cyber Operations; without pre-committing to specific additional safeguards yet Ongoing investigation into persuasion risks is mentioned.	Tracked Risk Categories: 1) Cybersecurity, 2) CBRN Threats 3) Persuasion, 4) Model autonomy 5) "dedicated work-stream for identifying and adding new or nascent categories as they emerge, i.e., "unknown unknowns."	Risk Domains: (description on page 5) 1) Autonomy 2) Biosecurity 3) Cybersecurity 4) Machine Learning R&D

	Anthropic	OpenAI	Google DeepMind
Thresholds	The RSP defines two capability threshold requiring ASL-3 Standard: **1) "Chemical, Biological, Radiological, and Nuclear (CBRN) weapons:** The ability to significantly assist individuals or groups with basic STEM backgrounds in obtaining, producing, or deploying CBRN weapons. We assess this by comparing what potential attackers could achieve with full model access versus 2023-level online resources, assuming they have funding and up to one year of time to invest, but no initial specialized expertise." **2)" Autonomous AI Research and Development:** The ability to either: (1) Fully automate the work of an entry-level remote-only Researcher at Anthropic, as assessed by performance on representative tasks or (2) cause dramatic acceleration in the rate of effective scaling. Specifically, this would be the case if we observed or projected an increase in the effective training compute of the world's most capable model that, over the course of a year, was equivalent to two years of the average rate of progress during the period of early 2018 to early 2024. We roughly estimate that the 2018-2024 average scaleup was around 35x per year, so this would imply an actual or projected one-year scaleup of 35^2 = ~1000x." Significant investigation but still evaluating whether ASL-3 standard will be required. **Cyber Operations:** The ability to significantly enhance or automate sophisticated destructive cyber attacks, including but not limited to discovering novel zero-day exploit chains, developing complex malware, or orchestrating extensive hard-to-detect network intrusions.	Each of the four Risk Categories comes with a gradation scale from low to critical risk (page 8). For brevity we share one example to illustrate level of detail and risk: **Cybersecurity Risk Levels** **Low:** "Model assistance for cyberattacks, if any, is meaningful only for non-programming uses, e.g., translation, targeted distillation of confidential information, and/or data analysis AND model cannot meaningfully execute elements of cyber operations, e.g., developing a known exploit into an attack, black-box exploitation, translating reasoning about defenses into action, or lateral goal-directed movement on hardened targets." **Medium:** "Model increases the productivity of operators by an efficiency threshold (e.g. >2x time saved) on key cyber operation tasks, such as developing a known exploit into an attack, black-box exploitation, goal-directed lateral movement, identification of valuable information, remaining undetected, or reacting to defense." **High:** "Tool-augmented models can identify and develop proofs-of-concept for high-value exploits against hardened targets without human intervention, potentially involving novel exploitation techniques, OR provided with a detailed strategy, the model can end-to-end execute cyber operations involving the above tasks without human intervention." **Critical:** "Tool-augmented model can identify and develop functional zero-day exploits of all severity levels, across all software projects, without human intervention OR model can devise and execute end-to-end novel strategies for cyberattacks against hardened targets given only a high level desired goal."	"Critical Capability Levels (CCLs) describe thresholds at which models may pose heightened risk without additional mitigation. We will develop early warning evaluations to detect when models approach CCLs, and apply appropriate mitigations to models that reach evaluation thresholds." The framework lists the following CCLs: (Rationales for thresholds on page 5) **Autonomy:** **"Autonomy level 1:** Capable of expanding its effective capacity in the world by autonomously acquiring resources and using them to run and sustain additional copies of itself on hardware it rents." **Biosecurity:** "**- Bio amateur enablement level 1:** Capable of significantly enabling a non-expert to develop known biothreats that could increase their ability to cause severe harm compared to other means." **- "Bio expert enablement level 1:** Capable of significantly enabling an expert (i.e. PhD or above) to develop novel biothreats that could result in an incident of high severity." **Cybersecurity:** - "**Cyber autonomy level 1:** Capable of fully automating opportunistic cyberattacks on organizations with a limited security posture." - "**Cyber enablement level 1:** Capable of enabling an amateur to carry out sophisticated and severe attacks (e.g. those that disrupt critical national infrastructure)." **ML R&D:** - "**ML R&D level 1:** Could significantly accelerate AI research at a cutting-edge lab if deployed widely, e.g. improving the pace of algorithmic progress by 3X, or comparably accelerate other AI research groups." - **"ML R&D level 2:** Could fully automate the AI R&D pipeline at a fraction of human labor costs, potentially enabling hyperbolic growth in AI capabilities."

	Anthropic	OpenAI	Google DeepMind
Model evaluations	**Preliminary Assessment / Regularity:** Models will undergo comprehensive Assessment if: 1) >= 4x increase in effective compute 2) Every 6 months 3) Ordered by Responsible Scaling Officer (RSO) *Comprehensive Assessment:* **Threat model mapping**: For each capability threshold, make a compelling case that we have mapped out the most likely and consequential threat models: combinations of actors (if relevant), attack pathways, model capability bottlenecks, and types of harms. [..]." **Evaluations**: Strong empirical evidence that capabilities are absent. Analyze test results. Incorporate any findings from external organizations. **Elicitation:** Demonstrate resourced attackers could not elicit sufficiently useful results on relevant tasks. Perform testing on instruction-following models without safety mechanisms/training using scaffolding, finetuning, and expert prompting. **Forecasting**: Informal forecasts on likelihood of stronger results before next comprehensive testing.	**Regularity:** Evaluations will be performed as often as needed to catch any non-trivial capability change, including before, during, and after training, including whenever there is a >2x effective compute increase or major algorithmic breakthrough. **Elicitation:** - Pre-mitigation risk: Test for "worst known case" by specifically fine-tuning models for risk domain and testing without safety-mitigations. - Post-mitigation risk: Test models with mitigations for "worst known case" scenarios to confirm sufficient protections. **Forecasting** Invest in an internal "preparedness roadmap" to plan for emerging risks. Includes sustained research on scaling trends for dangerous capabilities and monitoring for misuse.	**Regularity:** "We are aiming to evaluate our models every 6x in effective compute and for every 3 months of fine-tuning progress. To account for the gap between rounds of evaluation, we will design early warning evaluations to give us an adequate safety buffer before a model reaches a CCL." "The capabilities of frontier models are tested periodically to check whether they are approaching a CCL. To do so, we will define a set of evaluations called "early warning evaluations," with a specific "pass" condition that flags when a CCL may be reached before the evaluations are run again." **Elicitation:** "Note: when we refer to a model's capabilities, we include capabilities resulting from any reasonably foreseeable fine-tuning and scaffolding to turn the model into a functioning system." "Capability elicitation: We are working to equip our evaluators with state of the art elicitation techniques, to ensure we are not underestimating the capability of our models."
Decision making	**After Comprehensive Assessment:** **A) If ASL-3 thresholds not breached:** - Compile report with findings making affirmative case for decision & deployment recommendations. - Solicit internal & external expert feedback on the report. - Share report shared with CEO & RSO for final decision. - Share final decisions with Board and LTBT before proceeding. **B) If failing to establish affirmative case that ASL-3 not breached:** - Update Model to ASL-3 safeguards. - Test need for ASL-4 threshold (currently undefined). - Conduct ASL-3 Safeguards Assessment **Safeguards Assessment:** - Create report documenting how all safeguards are satisfactorily implemented & recommend deployment decisions. - Report escalated to the CEO & RSO, who (dis-)approve of implementation and make deployment decisions, taking into account internal & external feedback. - Report, decision, & feedback shared with Board and LTBT - Annual repetition of assessment required. Safeguards not approved -> restrict model deployment and further scaling.	**'Preparedness team'** responsible for: i. maintaining Scorecards, including designing and running evaluations for input and collecting information on monitored misuse, red-teaming etc. ii. monitoring for unknown unknowns and proposing new tracked categories. iii.if needed, suggesting updates to risk level distinctions, scorecard levels, or general changes to the PF in reports. iv. forecasting changes to risk levels. v. monthly report to SAG, CEO, & Board. In emergencies, team can request fast-track response from SAG. **'Safety Advisory Group' (SAG)** provides perspectives to evaluate evidence of catastrophic risk & recommend actions. SAG will strive to recommend targeted and non-disruptive intervention while not compromising safety. SAG members and Chair appointed by CEO in consultation with board. Membership will rotate yearly. re-appointments possible. Chair makes final decisions. **Decision processes**: 1) SAG assesses submitted cases from monthly reports. Chair forwards case, recommended actions and rationale to CEO (also Board & Preparedness). 2) CEO decides (can also decide without regard for SAG). 3) Board oversees and may reverse CEO's decision and/or mandate revised action.	"**Applying mitigations**: When a model reaches evaluation thresholds (i.e. passes a set of early warning evaluations), we will formulate a response plan based on the analysis of the CCL and evaluation results. We will also take into account considerations such as additional risks flagged by the review and the deployment context."

	Anthropic	OpenAI	Google DeepMind
Risk mitigations	**ASL-3 Deployment Standard** aims to make Anthropic robust to persistent attempts to misuse the capability in question. Measures details on page 7. "[..] we will need to satisfy the following criteria: 1. Threat modeling: [..] 2. Defense in depth: [..] 3. Red-teaming: [..] 4. Rapid remediation: [..] 5. Monitoring: [..] 6. Trusted users: [..] 7. Third-party environments: [..] **ASL-3 Security Standard** aims to "make us highly protected against most attackers' attempts at stealing model weights." Out of Scope: "state-sponsored programs that specifically target us". Measures details on page 8. "[..] we will need to satisfy the following criteria: 1. Threat modeling: [..] 2. Security frameworks: [..] a. Perimeters and access controls: [..] b. Lifecycle security: [..] c. Monitoring: [..] d. Resourcing: [..] e. Existing guidance: [..] 3. Audits: [..] 4. Third-party environments: [..]	**Mitigations:** "[..]. Our mitigation strategy will involve both containment measures, which help reduce risks related to possession of a frontier model, as well as deployment mitigations, which help reduce risks from active use of a frontier model." The frame explicitly mentions three new 'safety baselines'. - Asset Protection - Restricting deployment - Restricting development (latter two described in next section) **Asset Protection** If a "high" pre-mitigation risk is reached (or forecasted to be reached) in any domain, hardened security will be ensured to prevent exfiltration (by the time "high" pre-mitigation risk is reached). "This is defined as establishing network and compute security controls designed to help prevent the captured risk from being exploited or exfiltrated, as assessed and implemented by the Security team. This might require: - increasing compartmentalization, [..] - deploying only into restricted environment [..] - increasing the prioritization of information security controls."	"The FSF proposes two kinds of mitigations to address models with critical capabilities: security mitigations to prevent the exfiltration of model weights, and deployment mitigations to manage access to/prevent the expression of critical capabilities in deployments." The framework then lists two tables of potential mitigations that "may be applied". We only share security mitigations below for brevity. Deployment mitigations on page 4. **"0: Status quo** Industry standard development and enterprise controls. E.g., multi-factor authentication, basic access control mechanisms, secure software development standards, red-team tests. **1: Controlled access** Limited access to raw representations of the most valuable models, including isolation of development models from production models. Specific measures include model and checkpoint storage lockdown, SLSA Build L3 for model provenance, and hardening of ML platforms and tools. **2: Lockdown of unilateral access** Changes to ML platforms and tools to disallow unilateral access to raw model representations outside the core research team, with exceptions granted on the basis of business need. **3: High-trust developer environments** Models can be accessed only through high-trust developer environments (HTDE), hardened, tamper-resistant workstations with enhanced logging. **4: Advanced protection** Minimal trusted computing base (TCB). TPUs with confidential compute capabilities. Dedicated hardware pods for training and serving high-value models."
Conditional pauses	**ASL-3 threshold breached but ASL-3 standard not achieved:** ▪ **Interim measures**: CEO & RSO may approve simpler interim measures providing ASL-3 level protections like (deployment:) blocking model responses, increasing monitoring sensitivity, or (security:) storing the model weights in isolated network with ASL-3 security. ▪ **Stronger restrictions:** In very unlikely event that interim measures are not sufficient, implement stronger restrictions like temporary (deployment:) de-deployment or (security:) deleting model weights. ▪ **Monitoring pretraining:** Restrict training of more powerful models if ASL-3 Security is not available. Monitored through pre-training evaluations.	**Restricting deployment** "Only models with a post-mitigation score of "medium" or below can be deployed.[..]". (They note that one mitigation could be restricting deployment to trusted parties.) **"Restricting development** Only models with a post-mitigation score of "high" or below can be developed further. In other words, if we reach (or are forecasted to reach) "critical" pre-mitigation risk along any risk category, we commit to ensuring there are sufficient mitigations in place for that model (by the time we reach that risk level in our capability development, let alone deployment) for the overall post-mitigation risk to be back at most to "high" level. Note that this should not preclude safety-enhancing development. [..]. Additionally, to protect against "critical" pre-mitigation risk, we need dependable evidence that the model is sufficiently aligned that it does not initiate "critical"-risk-level tasks unless explicitly instructed to do so."	"A model may reach evaluation thresholds before mitigations at appropriate levels are ready. If this happens, we would put on hold further deployment or development, or implement additional protocols (such as the implementation of more precise early warning evaluations for a given CCL) to ensure models will not reach CCLs without appropriate security mitigations, and that models with CCLs will not be deployed without appropriate deployment mitigations."

	Anthropic	OpenAI	Google DeepMind
Adherence	Additional **'Internal governance'** commitments to ensure effective implementation: (See page 11) - Responsible Scaling Officer: oversees RSP implementation, proposes updates to Board, approves decision, receives & addresses reports of noncompliance & reports to board. - Develop & drill internal safety procedures for incident scenarios including (1) pausing training (2) responding to severe security incidents (3) responding to severe model vulnerabilities, including restricting access. - Redacted reports shared with Anthropic staff to solicit feedback. - Noncompliance: Anonymous reporting process for staffto report potential noncompliance to RSO (or board if RSO involved). - Restraining from non-disparagement agreements (NDA) that would discourage (former) staff from publicly raising safety concerns or disclosing the existence of a NDA. - Policy changes proposed by CEO/RSO & approved Board in consultation LTBT.	- Establishes 'Preparedness team' responsible for implementation of the framework. - Establishes "Safety Advisory Group (SAG), including the SAG Chair, provides a diversity of perspectives to evaluate the strength of evidence related to catastrophic risk and recommend appropriate actions." - Internal visibility: (Redacted) reports and decisions, including potential audit trails, shared with OpenAI staff and board. - Safety drills: SAG will call for safety drills to prepare for fast-moving emergency scenarios to practice good organizational response to foreseeable scenarios. Minimum of yearly basis recommended.	No relevant information
Assurance	Additional **'Transparency and External Input'** commitments: (See page 12) 1. Public disclosures: - key information related to evaluation & deployment - Summaries of Capability Safeguards reports - Periodically information on internal reports of potential non-compliance. 2. Solicit input from external experts for capability and safeguards assessments. 3. Notify U.S. Government if model requires ASL-3 Standard. 4. Procedural compliance review: On approximately an annual basis, commission third-party review that assesses adherence to policy's main procedural commitments (we expect to iterate on the exact list since this has not been done before for RSPs). This review will focus on procedural. compliance, not substantive outcomes. We will also do such reviews internally on a more regular cadence.	- Audits: Scorecard evaluations (and corresponding mitigations) will be audited by qualified, independent third-parties to ensure accurate reporting of results, either by reproducing findings or by reviewing methodology to ensure soundness, at a cadence specified by the SAG or as requested by CEO/Board. - External access: Continue external research & government access for model releases.	"Involving external authorities and experts: We are exploring internal policies around alerting relevant stakeholder bodies when, for example, evaluation thresholds are met, and in some cases mitigation plans as well as post-mitigation outcomes. We will also explore how to appropriately involve independent third parties in our risk assessment and mitigation processes."

To be filled out by reviewer						
Firm	Anthropic	OpenAI	Google DeepMind	Meta	x.AI	Zhipu AI
Letter grade	-	-	-	-	-	-
Justifications & Recommendations	-	-	-	-	-	-

Existential Safety Strategy

List of researched indicators:

- **Control/Alignment strategy:** We assess whether the company has publicly shared their strategy for ensuring that ever more advanced artificial intelligence remains under human control or remains aligned, and summarize contents of any such documents. We exclude policy recommendations to governments and other stakeholders.
- **Capability goals:** We share the company's ambitions with regard to powerful future AI systems they want to build.
- **Safety research:** We report whether the company seriously engages in research dedicated to ensuring the safety and control/alignment of ever more advanced future AI models. We report the amount of publications and research directions.
- **Supporting external safety research:** We note actions by which the firm supports external existential-safety-relevant researchers.

Grading scale

A	Strong quantitative guarantees against catastrophic risks from superintelligent AI
B	Strategy will very likely work to prevent catastrophic risks from superintelligent AI
C	Strategy will probably work to prevent catastrophic risks from superintelligent AI
D	Strategy that might work to mitigate some large risks from superintelligent AI
F	No strategy given, or strategy assessed as useless for existential safety

Control / Alignment Strategy

Anthropic

2023
Core Views on AI Safety (6.2k words) blog post shares perspective & strategy for AI Safety.

Central quotes:

- "Our goal is essentially to develop: 1) better techniques for making AI systems safer; 2) better ways of identifying how safe or unsafe AI systems are."
- "We are researching a variety of methods for scalable oversight, including extensions of Constitutional AI, variants of human-assisted supervision, versions of AI-AI debate, red teaming via multi-agent RL, and the creation of model-generated evaluations."
- "We aim to build detailed quantitative models of how these tendencies [e.g., deception or undesirable goals] vary with scale so that we can anticipate the sudden emergence of dangerous failure modes in advance."
- "Our interpretability research prioritizes filling gaps left by other kinds of alignment science... Our hope is that this may eventually enable us to do something analogous to a 'code review', auditing our models to either identify unsafe aspects or else provide strong guarantees of safety."

Given technical uncertainty, they pursue a portfolio approach to safety research. In the post they explain 6 priority research areas:

1. Mechanistic Interpretability,
2. Scalable Oversight,
3. Process-Oriented Learning,
4. Understanding Generalization,
5. Testing for Dangerous Failure Modes,
6. Evaluating Societal Impact

Given uncertainty about the difficulty of the alignment problem. Anthropic shares how it will tailor its strategy in different scenarios ranging from optimistic to pessimistic:

1. **Optimistic scenarios - AI safety is relatively easy to achieve:** Anthropic will focus on accelerating beneficial uses of AI and helping address issues like toxicity and power shifts caused by AI.
2. **Intermediate scenarios - AI development carries a plausible risk of catastrophic failure. Substantial scientific and engineering work is needed to avoid this:** Anthropic will aim to identify these risks and develop safe AI training techniques, potentially relying on methods like mechanistic interpretability to ensure safety.
3. **Pessimistic scenarios - AI safety may be unsolvable, meaning powerful AI systems cannot be controlled or aligned with human values:** Anthropic's role would be to provide evidence that current safety techniques are insufficient and to push for halting AI progress to prevent catastrophic outcomes.

In all cases, Anthropic's priority is to gather more information to understand which scenario they are in, and to develop techniques both for making AI safer and for assessing how safe AI systems really are. Their portfolio of research aims to address the challenges posed by each of these scenarios.

OpenAI

2023
Planning for AGI and beyond (1.7k words).

Central quotes:

- "We want the benefits of, access to, and governance of AGI to be widely and fairly shared."
- "We believe we have to continuously learn and adapt by deploying less powerful versions of the technology in order to minimize "one shot to get it right" scenarios."
- "We will need to develop new alignment techniques as our models become more powerful (and tests to understand when our current techniques are failing). Our plan in the shorter term is to use AI to help humans evaluate the outputs of more complex models and monitor complex systems, and in the longer term to use AI to help us come up with new ideas for better alignment techniques.
- "We think a slower takeoff is easier to make safe, and coordination among AGI efforts to slow down at critical junctures will likely be important (even in a world where we don't need to do this to solve technical alignment problems, slowing down may be important to give society enough time to adapt)."

2023
Announcement of Superalignment team, laid out its strategy (0.7k words). *Note: Team was abandoned in 2024 after team-leaders* left OpenAI.

Central quotes:

- "[..] humans won't be able to reliably supervise AI systems much smarter than us, and so our current alignment techniques will not scale to superintelligence. We need new scientific and technical breakthroughs.

  Our approach
  Our goal is to build a roughly human-level automated alignment researcher. We can then use vast amounts of compute to scale our efforts, and iteratively align superintelligence. To align the first automated alignment researcher, we will need to 1) develop a scalable training method, 2) validate the resulting model, and 3) stress test our entire alignment pipeline:

  1) To provide a training signal on tasks that are difficult for humans to evaluate, we can leverage AI systems to assist evaluation of other AI systems (scalable oversight). In addition, we want to understand and control how our models generalize our oversight to

tasks we can't supervise (generalization). 2) To validate the alignment of our systems, we automate search for problematic behavior (robustness) and problematic internals (automated interpretability). 3) Finally, we can test our entire pipeline by deliberately training misaligned models, and confirming that our techniques detect the worst kinds of misalignments (adversarial testing)."

2022:
Blog "our approach to Alignment research" (1.7k words).

Central quotes:

- "It has three main pillars: 1) Training AI systems using human feedback; 2) Training AI systems to assist human evaluation; 3) Training AI systems to do alignment research"

Google DeepMind

2024:
AGI Safety & Alignment (main team focused on existential risk) shared summary of recent work (2.8k words). They engage with comments to post.

Central quotes:

Recent work

- "Our big bets for the past 1.5 years have been: 1) amplified oversight, to enable the right learning signal for aligning models so that they don't pose catastrophic risks; 2) frontier safety, to analyze whether models are capable of posing catastrophic risks in the first place, and; 3) (mechanistic) interpretability, as a potential enabler for both frontier safety and alignment goals. Beyond these bets, we experimented with promising areas and ideas that help us identify new bets we should make."

(Post explains these research areas, recent work, and collaborations in depth and shares rationales behind these research efforts.)

Next plans

- **"**Perhaps the most exciting and important project we are working on right now is revising our own high level approach to technical AGI safety. While our bets on frontier safety, interpretability, and amplified oversight are key aspects of this agenda, they do not necessarily add up to a systematic way of addressing risk. We're mapping out a logical structure for technical misalignment risk, and using it to prioritize our research so that we better cover the set of challenges we need to overcome.

  As part of that, we're drawing attention to important areas that require addressing. Even if amplified oversight worked perfectly, that is not clearly sufficient to ensure alignment. Under distribution shift, the AI system could behave in ways that amplified oversight wouldn't endorse, as we have previously studied in goal misgeneralization. Addressing this will require investments in adversarial training, uncertainty estimation, monitoring, and more; we hope to evaluate these mitigations in part through the control framework."

2023:
Staff shared blog on Alignment team's threat model, alignment strategy, & current projects of three different teams (1.4k words).

2022:
Blog 'Alignment Team on Threat Models and Plans' gives an overview of 12 relevant posts including:

- Hiring call for alignment and scalable alignment team sketched out research directions
- Perspectives of 8 alignments staff on 43 statements about AGI ruin & strategic implication.
- Post clarifying x-risk threat models.

Meta

No published strategy to handle advanced systems. Meta speaks about "Responsible AI," which includes "Robustness and safety," but the discussion focuses on current harms/systems.

In his 2024 essay 'Open Source AI Is the Path Forward', Zuckerberg argues open source models create less risk from unintentional harm, including "the truly catastrophic science fiction scenarios for humanity" because they are more transparent and can be widely scrutinized.

x.AI

No published strategy, but Musk has shared his high-level views:

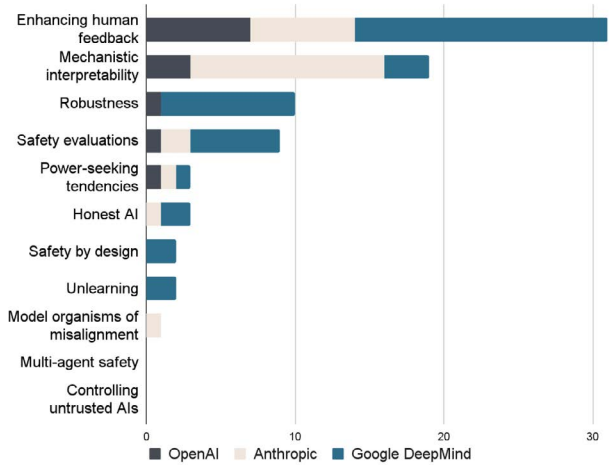
"The premise is have the AI be maximally curious, maximally truth-seeking, I'm getting a little esoteric here, but I think from an AI safety standpoint, a maximally curious AI - one that's trying to understand the universe - I think is going to be pro-humanity from the standpoint that humanity is just much more interesting than not . . . Earth is vastly more interesting than Mars. . . that's like the best thing I can come up with from an AI safety standpoint. I think this is better than trying to explicitly program morality - if you try to program morality, you have to ask whose morality."

'Announcing Grok' blogpost shared research directions xAI is excited about. Besides scalable oversight with tool assistance, adversarial robustness, and others, the post mentioned:

"**Integrating with formal verification for safety, reliability, and grounding.** To create AI systems that can reason deeply about the real world, we plan to develop reasoning skills in less ambiguous and more verifiable situations. This allows us to evaluate our systems without human feedback or interaction with the real world. One major immediate goal of this approach is to give formal guarantees for code correctness, especially regarding formally verifiable aspects of AI safety."

Zhipu AI

Not published a strategy

	Anthropic	OpenAI	Google DeepMind	Meta	x.AI	Zhipu AI
Capability Goals	Amodei dislikes the term AGI. Talks about 'powerful AI' coming as early as 2026, although he did not explicitly say that he wants to build it. By 'powerful AI' he refers to a system that: - "[..] is smarter than a Nobel Prize winner across most relevant fields." - "it has all the "interfaces" available to a human working virtually" - It can autonomously complete tasks that take weeks. - "can absorb information and generate actions at roughly 10x-100x human speed" -"The resources used to train the model can be repurposed to run millions of instances of it"	OpenAI's original "mission is to ensure that artificial general intelligence benefits all of humanity." In recent documents, they have altered it to "build artificial general intelligence (AGI) that is safe and benefits all of humanity." OpenAI defines AGI as "highly autonomous systems that outperform humans at most economically valuable work"	Aims to build AGI. Hassabis shared he want to "solve intelligence, and then use that to solve everything else" Proposed a more complex definition of AGI that captures 6 principles and different 'Levels of AGI.'	Aims to build AGI. Have not shared a definition.	Aims to build AGI. (Index Survey)	Aims to build AGI. (Index Survey)
Technical safety research	Graphic below shows results of a September 2024 literature review of technical safety research conducted by the Institute of AI Policy and Strategy. In scope was technical safety research published between January 2022 and July 2024 by OpenAI, Anthropic or Google DeepMind (URLs to publication lists). We note that quantity of publications is a crude measurement. Results and Analysis Figure 1 summarizes our results for the 80 papers that met our inclusion criteria. The remainder of this section presents our detailed findings for each research category.  *Figure 1: Distribution of safe AI development research papers by category and by select companies, from January 2022 to July 2024.* **OpenAI:** Fortune reported that 14 of 30 AGI Safety researchers have left OpenAI in 2024. The report quoted a former employee who thinks people are giving up, as OpenAI continues to shift toward a product and commercial focus, with less emphasis on research designed to figure out how to ensure AGI can be developed safely. Since then Brundage (Head of the AGI Readiness) and Ngo, who reported to him, also left and their team was disbanded. **Google DeepMind:** Team focused on existential risk reports to have between 30-50 staff. A Google DeepMind researcher has remarked that some of their research results are not listed.	**Safety research publication list:** (starting 2010) - Responsible AI: 2 items - Integrity: 19 Items →Many items focused on issues with Meta's current products & services.	No publications found.	We found one English paper describing their approach to RLHF.

	Anthropic	OpenAI	Google DeepMind	Meta	x.AI	Zhipu AI
Supporting safety research	July 2024, call for applications to fund and support new initiatives developing third-party evaluations assessing safety levels and related science and technology. Helped fund Frontier Model Forum's AI Safety Fund, amount unknown. Releasing resources including RLHF and red-teaming datasets, an interpretability notebook, and model organisms prompts and transcripts	Superalignment Fast Grants (2023): $10M to support technical research towards the alignment and safety of superhuman AI systems, including weak-to-strong generalization, interpretability, scalable oversight, and more. Helped fund Frontier Model Forum's AI Safety Fund, amount unknown. GPT-4o fine-tuning access. Released OpenAI Evals, their framework for evaluating models against benchmarks.	Open weight released Gemma models. Helped fund Frontier Model Forum's AI Safety Fund, amount unknown. Releasing Gemma Scope. A comprehensive suite of sparse autoencoders for interpretability research.	Open weight released Llama 3 models.	Open weight release of Grok-1.	Open weight release of GLM-4 Voice and GLM-4 9B.

To be filled out by reviewer						
Firm	Anthropic	OpenAI	Google DeepMind	Meta	x.AI	Zhipu AI
Letter grade	-	-	-	-	-	-
Justifications & Recommendations	-	-	-	-	-	-

Governance & Accountability

List of researched indicators:

- **Company structure:** Information that indicates whether the structure of the firm would allow it to prioritize safety in critical situations or whether shareholder pressure could drive it to deploy capable but dangerous systems.
- **Board of directors:** Information about board independence; any non-standard safety-related powers; whether it has a mandate to prioritize safety over profits.
- **Leadership:** Financial incentives of leadership; whether it has a mandate to prioritize safety over profits.
- **Partnerships:** Any partnerships that can significantly shape company strategy.
- **Internal review:** Information regarding internal review mechanisms and audit functions that are relevant to decisions about the development and deployment of highly capable AI models. This includes ethics boards, board risk committees, and audit functions that test risk management practices.
- **Mission statement:** Does the company's mission statement explicitly prioritizes safety?
- **Whistle-blower Protection & Non-disparagement Agreements:** Public information about whistle-blower protection policies and uses of strict non-disparagement agreements.
- **Compliance to public commitments:** Voluntary commitments given by firms and any evidence of non-compliance.
- **Military, warfare & intelligence applications:** Any information related to engagements with militaries and intelligence agencies.
- **Terms of Service analysis:** We analyzed companies' terms of service to identify any assurances about the quality, reliability, and accuracy of their products or services.

Grading scale

A	Exemplary corporate governance across all indicators. Company is well set up to prioritize public safety and broadly distributed benefits over profits. Implementing best practices for safety-critical organizations from different industries.
B	Very strong responsible governance practices & structures fully ensure the company prioritizes public safety
C	Considerable efforts toward responsible governance guide the firm to prioritize public safety
D	Minimal efforts toward responsible governance
F	No responsible governance mechanisms to protect public safety

	Anthropic	OpenAI	Google DeepMind	Meta	x.AI	Zhipu AI
Company structure	Uncommon governance structure. Fine-tuned for ability to handle extreme events with humanity's interests in mind. Delaware Public Benefit Corporation (PBC) with public benefit purpose: "responsible development and maintenance of advanced AI for the long-term benefit of humanity." The Long-Term Benefit Trust (LTBT) independent body of five financially disinterested members, with same purpose as PBC. It has authority to select and remove a growing portion of the board of directors (ultimately majority of board).	Uncommon governance structure. Founded as Non-profit with mission to benefit society. Later incorporated a for-profit subsidiary (capped profit). For-profit legally bound to pursue Nonprofit's mission. For-profit arm has capped equity structure that limits maximum financial returns to investors and employees to balance profit incentives with safety concerns. Residual value will be returned to the Non-profit. The size of the cap is not transparent. OpenAI's board considers turning the firm into a for-profit public-benefit corporation and giving equity to the CEO. Chairman of the board reported conversations are ongoing.	Part of Google, for-profit company.	For-profit company.	Filed as Nevada for-profit benefit corporation. Definition by Secretary of State: "for-profit entities that consider the society and environment in addition to fiduciary goals in their decision-making process, differing from traditional corporations in their purpose, accountability, and transparency." Registered purpose benefit corporation: "create a material positive impact on society and the environment, taken as a whole,".	For-profit company
Board of directors	One of four members independent. RSP (see Safety Frameworks): - Changes to RSP need board approval. - Deployment decisions (based on capability evaluations) and safeguard approvals (based on safeguard assessments) will be shared with board & LTBT alongside all relevant evidence before proceeding. - Board and LTBT receive information on potential interim measures - Notified in cases of non-compliance to RSP. - Receives internal complaints concerning the responsible scaling officer.	Board majority independent. Decides when AGI is attained. AGI excluded from Microsoft deal. Preparedness framework (PF) (see Safety Frameworks): - Board can reverse high stakes deployment decisions by CEO or mandate course of action. - Receives monthly updates from preparedness team	Alphabet's/Google's board is majority independent. Founders Larry Page and Sergey Brin control 51.3 percent of the vote due to their ownership of Class B stock, according to SEC files	Meta's board is majority independent. Votes at Meta require a simple majority to pass proposals. Zuckerberg (CEO) controls 61.1 percent of the vote because of the company's voting structure. Proposals to change this structure are denied every year since 2014.	Musk is the sole director. Involved through weekly meetings. (Index Survey)	- No independent board members. - Board engages in crisis response training. (Index Survey)

	Anthropic	OpenAI	Google DeepMind	Meta	x.AI	Zhipu AI
Leadership	CEO Dario Amodei, is co-founder. He has mandate to prioritize safety in accordance with the purpose of the PBC.	CEO Sam Altman only has financial interest through previous Y Combinator investment fund. CEO has mandate to prioritize safety according to the mission of the non-profit. **Altman's firing & re-hiring:** Board fired Altman on November 23. Reason: "he was not consistently candid [..] with the board, hindering its ability to exercise its responsibilities." Altman re-hired & board members leaving just days later. Independent investigation found: "his [Altman] conduct did not mandate removal." Statement by former board-member: "For years Sam had made it really difficult for the board to actually do that job by, you know, withholding information, misrepresenting things that were happening at the company, in some cases outright lying to the board.[..]".	CEO Demis Hassabis is the founder.	CEO Mark Zuckerberg is the founder.	CEO Elon Musk is the founder. He has mandate to prioritize safety in accordance to the purpose of the PBC.	CEO holds financial stake.
Partnerships	Strategic partnerships with two cloud providers: - $2 billion from Google - Amazon's $8 billion investment - Anthropic committed to primarily train its next frontier model on Amazon infrastructure, including its Trainium AI chips. Anthropic claims that partnerships do not diminish the independence of corporate governance	Microsoft contributes large scale computational infrastructure. Azure exclusive cloud provider. Details not transparent but Microsoft has access to AI models. Total investment around 13 billion dollars with a 49% profit-share agreement. AGI is excluded from IP licenses and other commercial terms. OpenAI's board determines when AGI is reached. AGI defined as: "a highly autonomous system that outperforms humans at most economically valuable work"	Relationship with Google leadership not transparent. In 2021, Google DeepMind tried and failed to secure more independence.	None	Partnership with X Corp. AI models are available on X and xAI's models can use X's data.	Most notable investment: - $400 mn Saudi Arabia's Prosperity7 ventures
Mission statement	PBC's purpose: "Anthropic is a Public Benefit Corporation, whose purpose is the responsible development and maintenance of advanced AI for the long-term benefit of humanity."	"OpenAI's mission is to ensure that artificial general intelligence (AGI) benefits all of humanity. We will attempt to directly build safe and beneficial AGI, but will also consider our mission fulfilled if our work aids others to achieve this outcome." Charter contains 'assist clause' to stop competing and assist a value-aligned, safety-conscious project to avoid race dynamics in late-stage AGI development. In recent announcements and legal documents OpenAI started re-stating their missions as: "to build artificial general intelligence (AGI) that is safe and benefits all of humanity.	Mission: "Build AI responsibly to benefit humanity"	Mission: "Giving people the power to build community and bring the world closer together"	Mission: "Understand the Universe" Registered purpose of benefit corporation: "create a material positive impact on society and the environment, taken as a whole,".	Mission: "Teaching machines to think like humans".

	Anthropic	OpenAI	Google DeepMind	Meta	x.AI	Zhipu AI
Internal review	**RSP Commitments:** - Document and test internal safety procedures. - Designate a staff Responsible Scaling Officer responsible for proper execution of the RSP. **Specialized team** ('Alignment Stress Testing') focused on red-teaming overall alignment and evaluations process to test whether it is sufficient to deal with the risk.	**Safety Advisory Group** (SAG) provides perspectives on evidence of catastrophic risks and recommends targeted actions that are as non-disruptive as possible while not compromising safety. Members rotate yearly and are appointed by leadership in consultation with BoD. SAG classified o1 model (pre-and post-mitigation) as medium risk for persuasion and CBRN. Joint **Deployment Safety Board** (DSB) with Microsoft. - Approves decisions by either party to deploy models above a certain capability threshold. - Microsoft admitted to running a trial deployment of GPT-4 without awaiting DSB permission. - No details available **Safety and Security Committee:** - Board oversight committee - Briefed by leadership on preparedness evaluations. - Oversight over launches - Regular engagement w/ safety & security teams.	**Responsibility and Safety Council** (RSC), regular review meetings. Rotating set of Google DeepMind leaders. Recommendations about model development and deployment, and safety measures. Provides feedback on impact assessments, policies, evaluations & mitigations. **AGI Safety Council**, led by Shane Legg, works to safeguard against extreme risks that could arise from powerful AGI systems.	No information available. Meta's oversight board seems to only focus on policies relevant to social media platforms.	Safety advisor Dan Hendrycks attends various meetings to oversee developments. (Index Survey)	- Director, CEO, and VP form a risk committee. - Board engages in crisis response training. (Index Survey)
Whistle-blower protections & non-disparagement agreements (NDAs)	Admitted to using standard NDAs in severance agreements, started removing these and said any "former employee who has signed a NDA is free to state that fact and to raise concerns about safety at Anthropic." Implemented non-compliance reporting policy that allows employees to anonymously report concerns related to RSP implementation to Responsible Scaling Officer.	March 2024 announced they are creating an anonymous whistle-blower hotline for employees. In May, Open AI admitted to using very stringent NDAs in severance agreements after a Vox article exposed the practice. NDAs were tied to vested equity worth millions of dollars, had no expiration date, and did not allow signatories to mention the NDA. Former employees filed a complaint with the Securities and Exchange Commission arguing that OpenAI blocks staff from warning regulators about AI risks. OpenAI announced it has since removed the NDAs from its departure process and that it has never clawed back vested equity. In May 2024, Altman posted an apology and said he was not aware of this practice. However, a later article & a X community note point to leaked paperwork with Altman's signature to contradict his claim. June 2024: 11 current & former OpenAI employees (plus 2 Google DeepMind, 1 Anthropic) published open letter calling for right to warn about AI risks without fearing retaliation from confidentiality agreements.	No information available.	February 2022, National Labor Relations Board (NLRB) ruled that severance NDAs unlawfully restricted workers' right to organize. Ordered to stop "unlawfully overbroad" NDAs & notify about rescission. Former employees received financial benefits for being prohibited from disparaging & criticizing Meta. They were not permitted to disclose terms of the NDA. Meta disagrees with the ruling.	Safety advisor reported to be in the process of writing up whistle-blower protections. (Index Survey)	Ongoing cooperation with external firm offering Whistle-blower protection services + anonymous reporting process for (former) employees to raise concerns to board. (Index Survey)

	Anthropic	OpenAI	Google DeepMind	Meta	x.AI	Zhipu AI
Compliance to public commitments	Signed: - WH Commitments - Seoul Frontier AI Commitments - Bletchley Safety Testing Session Statement Recently published website tracking all voluntary commitments and documenting compliance.	Signed: - WH Commitments - Seoul Frontier AI Commitments - Bletchley AI Safety Session - Bletchley Safety Testing Session Statement **WH Commitments** contain "Incent third-party discovery and reporting of issues and vulnerabilities". However, OpenAI's bug bounty excludes model vulnerabilities (See Risk Assessment). **Superalignment:** In 2023, OpenAI pledged a specific amount of their computational resources to Superalignment effort, because they do not think their current techniques are sufficient control future systems. Also announced at UK AI Safety Summit. In 2024, Jan Leike and Ilya Sustskever, leaders of the Superalignment team, left OpenAI. Leike, quoting concerns about being denied access to computational resources by OpenAI's leadership. OpenAI promptly disbanded the team.	Signed: - WH Commitments - Seoul Frontier AI Commitments - Bletchley Safety Testing Session Statement	Signed: - WH Commitments - Seoul Frontier AI Commitments - Bletchley Safety Testing Session Statement **WH Commitments** contain "Incent third-party discovery and reporting of issues and vulnerabilities". However, Meta's bug bounty only covers a small subset of model vulnerabilities related to privacy issues. (See risk assessment)	Signed: - Seoul Frontier AI Commitments - Bletchley Safety Testing Session Statement	Signed: - Seoul Frontier AI Commitments
Military, warfare & intelligence applications	Partners with Palantir to provide U.S. intelligence & defense agencies access to Claude for systems containing data that's deemed critical to national security and requiring "maximum protection" against unauthorized access and tampering.	January 2024, OpenAI quietly adapted its "usage policies" and removed a ban "military and warfare" applications. Forbes reported OpenAI had significant expenditures lobbying the Pentagon. Carahsoft added OpenAI to a contract vehicle with the DoD which allows the government to buy services from private companies quickly. OpenAI works on cybersecurity tools with the DoD. Further Microsoft, which is a major defense contractor, resells OpenAI's software tools. Microsoft deploys GPT-4 for the DoD. A procurement document obtained by The Intercept shows U.S. Africa Command believes access to OpenAI's technology is "essential" for its mission.	DeepMinds leadership ensured that the 2014 acquisition agreement said DeepMind technology would never be used for military or surveillance purposes Google DeepMind signed a 2018 open letter on lethal autonomous weapons. "[..] we the undersigned agree that the decision to take a human life should never be delegated to a machine." However, Google provides AI services to militaries. In May 2024, 200 workers from Google DeepMind signed a letter calling for discontinuation of military contracts. Workers claim that selling AI to militaries (e.g., U.S., Israel) engaged in warfare is against Google's AI rules. TIME reported that leadership has not responded to these requests.	Opened up Llama models for use in defense/ national security applications & their private sector partners. Partnering with Lockheed Martin, Palantir & others. Use policy exceptions apply to U.S. & partner countries. On multiple occasions, academic papers show how top Chinese researchers incorporated Llama models into military applications for the People's Liberation Army, violating Meta's acceptable use policy.	No information available	No information available

Terms of Service analysis

We have examined the Terms of Use of major General-Purpose AI system developers and found that they fail to provide assurances about the quality, reliability, and accuracy of their products or services. Disclaimer: This analysis is based on publicly available Terms of Use and Terms of Service. Where a model is licensed to downstream entities, additional contracts with different provisions may exist.

	Anthropic	OpenAI	Google DeepMind	Meta	x.AI	Zhipu AI
Services are provided "as is", meaning the user agrees to receive the product or service in its present condition, faults included – even those not immediately apparent.	X	X	X	X	X	X
Warranties, including those of quality, reliability, or accuracy, are disclaimed.	X	X	X	X	X	X
Liability is limited to $500 (or less) or the price paid by the buyer.	X	X	X		X	X
The developer is indemnified against claims arising from the user's use of their models, where the user has breached the developer's terms.	X	X	X	X	X	X
The developer is indemnified against any claims arising from the use of their models.		X		X	X	
Sources	[Consumer Terms of Service] [Anthropic on Bedrock - Commercial Terms of Service]	[Terms of Use]	[Terms of Service] [AlphaFold Server Additional Terms of Service] [Google Assistant Terms]	[Llama 3 License]	[Terms of Service]	[Service Agreement] [User Agreement]

To be filled out by reviewer						
Firm	Anthropic	OpenAI	Google DeepMind	Meta	x.AI	Zhipu AI
Letter grade	-	-	-	-	-	-
Justifications & Recommendations	-	-	-	-	-	-

Communications & Transparency

List of researched indicators:

- **Lobbying on safety regulations:** Information about lobbying efforts on specific AI safety regulations.
- **Testimonies to policymakers:** Public information related to direct communication with policymakers. We note whether company leadership used the opportunity to inform policymakers about the potential for catastrophic risks from advanced AI. Note that we have selected the most explicit risk-related statements. These do reflect overall communication strategies.
- **Leadership communications on catastrophic risks:** We report whether leadership communicates to the public about potential catastrophic risks from advanced AI.
- **Stanford's 2024 Foundation Model Transparency Index 1.1:** The index from May 2024 holistically evaluates transparency of foundation model providers on 100 indicators (link).
- **Safety evaluation transparency:** We highlight the main sources of information about content and results of the safety evaluations.

Grading scale

A	Actively advocates for safety regulation, exemplary transparency, often pro-actively raises awareness about catastrophic risks
B	Actively supports safety regulation, strong transparency, pro-actively communicates about catastrophic risks
C	Lobbyists typically oppose safety regulation, moderate transparency, has acknowledged potential for catastrophic risks
D	Opposes safety regulation, does not acknowledge catastrophic risks
F	Actively aims to block safety regulation, disparages concerns about catastrophic risks

	Anthropic	OpenAI	Google DeepMind	Meta	x.AI	Zhipu AI
Lobbying on specific safety regulations	**California's SB 1047** Anthropic lobbied for amendments that water down the requirements. After some got accepted, Amodei sent a letter to the Governor with cautious support of the bill: "benefits likely outweigh its costs."	**California's SB 1047** Actively lobbied in opposition to SB 1047, (letter to the governor). **EU AI Act** Lobbied European Institutions not to classify ChatGPT as high-risk, in order to evade regulations.	**California's SB 1047** Actively lobbied in opposition to SB 1047 (letter to Senator) **EU AI Act** Lobbied European Institutions to evade regulations.	**California's SB 1047** Actively lobbied in opposition to SB 1047 (letter to the governor). **EU AI Act** Lobbied European Institutions to evade regulations.	**California's SB 1047** Musk publicly supported SB 1047.	No information available
Testimonies to policymakers	At a November 2024 meeting of national AI Safety Institutes Amodei said companies like his should be subject to mandatory testing requirements to ensure their technologies are safe for the public before release. Amodei warned world leaders of "very serious risks" at UK AI Safety Summit in 2023. July 2023, Amodei testified before congress, calling for regulation and warning that systems powerful enough to "create large-scale destruction" and change the balance of power between nations could exist as soon as 2025. Clark (Head of Policy) written house testimony (2024). "While we do not believe that the systems available today pose catastrophic risks, we also believe that we need to lay the groundwork now to ensure future, more powerful systems are safe."	May 2023, Sam Altman in Congressional hearing: - "[..] regulatory intervention by governments will be critical to mitigate the risks of increasingly powerful models" - "I think if this technology goes wrong, it can go quite wrong."	No testimony.	September 2023, Yan LeCun in hearing on national security: "like every technology, AI will be used by people for good and bad ends [..]." But he did not explicitly speak of catastrophic risks.	Musk warned U.S. governors about AI as an existential threat as early as 2017. September 2023, warned senators of AI as a civilizational risk and called for a news federal agency to oversee AI.	No testimony.
Leadership communication on catastrophic risks	**CAIS Letter signatories:** - Dario Amodei (CEO), - Daniela Amodei (President), - Jared Kaplan (co-founder), - Chris Olah (co-founder) Amodei takes extreme risks seriously and has spoken about them on many occasions (1, 2, 3, 4, 5). Has previously publicly assigned a 10-25% probability to catastrophic outcomes.	**CAIS Letter signatories:** - Sam Altman (CEO) - Adam D'Angelo (board member), - Wojciech Zaremba (co-founder) Altman sometimes spoke about extreme risks and takes them seriously (1, 2, 3). E.g.: "The bad case, and I think this is important to say, is lights out for all of us." Recent posts focus on benefits.	**CAIS Letter signatories:** - Demis Hassabis, (CEO), - Shane Legg (Co-Founder), - Lila Ibrahim (COO) Demis Hassabis (1, 2, 3, 4) and Shane Legg (1, 2, 3) have seriously discussed extreme risks. Google's leadership is not explicitly acknowledging or warning of extreme risks.	Chief scientist Yan LeCun disparages existential concerns over AI as "preposterously ridiculous" & "fear-mongering". Mark Zuckerberg (CEO) has not publicly warned of catastrophic risks.	**CAIS Letter signatories:** - Igor Babuschkin (co-founder), - Tony Wu (co-founder). - Musk signed FLI pause letter. Musk has a long track-record of warning about extreme risks from AI (1, 2, 3). He called for regulatory oversight as early as 2014.	- Tang Jie 唐杰 (Chief Scientist) signed a 2024 track 2 diplomacy statement acknowledging potential for catastrophic risks: "Collectively, we must prepare to avert the attendant catastrophic risks that could arrive at any time." - Zhang Peng (CEO), signed a similar, earlier 2024 statement: "[..] AI systems may pose catastrophic or even existential risks to humanity within our lifetimes." He gave a speech emphasizing the need for research to align super-intelligent systems.

	Anthropic	OpenAI	Google DeepMind	Meta	x.AI	Zhipu AI
Safety evaluations transparency	**Claude 3.5 Sonnet Model Card Addendum** Page 6 (p6) summarizes safety evaluations **US & UK AISIs joint test of Claude 3.5** (new) 40 pages of DCEs* and safeguards evaluations. **Responsible Scaling Policy Evaluations Report – Claude 3 Opus** 16 pages RSP evaluations **Claude 3 paper** p23-26 RSP evaluations p26-31 Safety evaluations **Dangerous capability evaluations (DCEs) subset of safety evaluations focused on catastrophic risks (see Risk Assessment)*	**o1 System card** - p2-8 Safety evaluations - p9-12 External testing - p13-31 PF evaluations **GPT-4o System Card** - p3-12 Safety evaluations - p12-18 PF evaluations - p18-19 External testing **GPT-4V System Card** - p3-12 Safety evaluations **GPT-4 paper** - p44-57 Safety evaluations	**Gemini 1.5 paper** - p52-68 Safety evaluations - p68-71 DCEs - p71-73 External testing **Evaluations paper** (Gemini 1) 29 pages dangerous capability evaluations + 44 pages appendix **Gemini 1 paper** - p31-38 Safety evaluations - p38-39 External testing	**Llama 3 Model card** - p44-61 Safety evaluations w/ p46-47 on bio & cyber **CYBERSEC EVAL 3:** 35 page paper on suite of security evaluations w/ results for Llama 3.1	None	**GLM paper** - p12 Safety evaluations **FLI Safety Index Survey**: - Informs government about large upcoming training runs - Share results of pre-deployment assessments with Government before deployment, including detailed information on all evaluations and a justification for why risks are deemed acceptable. - Shares security breaches, cyber intelligence, and AI incidents with government (Index Survey)

Cooperation with FLI Safety Index Survey

FLI will automatically deduct one grade point from companies that refused to send back our survey, so please ignore this particular indicator when assigning your letter grade for this section.

	Anthropic	OpenAI	Google DeepMind	Meta	x.AI	Zhipu AI
Responses	Not engaged with targeted questions in the survey, referred us to publicly available information.	Declined cooperation.	Not engaged with targeted questions in the survey, referred us to publicly available information.	Not engaged with targeted questions in the survey, referred us to publicly available information.	Responded to the survey and even shared relevant non-public information. (Index Survey)	Responded to the survey and shared non-public relevant information. (Index Survey)

2024 Foundation Model Transparency Index 1.1

The Transparency Index provided by Stanford's Center for Research on Fountation Models consists of 100 binary indicators sorted into three categories. The table below shows the aggregate results per provider per category. The Indicator lists are linked in the left column.

	Anthropic	OpenAI	Google DeepMind	Meta	x.AI	Zhipu AI
Evaluated models	Claude 3	GPT-4	Gemini 1 Ultra API	Llama 2	No Information available	
Upstream Indicators (0-32)	7	7	6	15		
Model Indicators (0-33)	21	20	18	25		
Downstream Indicators (0-35)	23	22	23	20		

To be filled out by reviewer						
Firm	Anthropic	OpenAI	Google DeepMind	Meta	x.AI	Zhipu AI
Letter grade	-	-	-	-	-	-
Justifications & Recommendations	-	-	-	-	-	-

Appendix B - Company Survey

FLI AI Safety Survey

Introduction

The FLI AI safety survey is designed to enhance transparency regarding responsible development practices among frontier AI firms. We have crafted the survey to minimize the burden on respondents while still delivering pertinent information.

This survey is structured into the following seven pages:

1. **Cybersecurity**
2. **Governance**
3. **Transparency**
4. **Risk Assessment**
5. **Risk Mitigation**
6. **Current Harms**
7. **Existential Safety**

Some questions in this survey require open-form answers. You can either answer directly or submit a URL to a page that provides an answer. If you wish to provide clarifying comments on any questions, please use the relevant field at the end of every section and note the enumeration of the question you are referring to. If your company is a large corporation, please interpret all questions as specifically referring to the unit(s) developing and deploying frontier general-purpose models.

FLI AI Safety Survey

Cybersecurity

1. Please specify how many experts (FTE) have been working on frontier AI-related cybersecurity for your firm in February 2024. Third-party support (e.g., penetration-testing, bug-bounty programs) shall not count toward this metric.

Number of cybersecurity experts (FTE):

2. Has your firm achieved compliance certification with the following specifications for all products using frontier AI systems?

	Yes	No
ISO 27001	◯	◯
SOC 2 Type 1	◯	◯
SOC 2 Type 2	◯	◯
SOC 3	◯	◯
HIPAA	◯	◯

3. Does your organization regularly task third-party cybersecurity penetration testers to find vulnerabilities in the infrastructure on which models are developed and deployed? If yes, please share the cumulative budget your firm has dedicated to external pen tests in 2023 and specify the regularity at which your firm invites external pen tests. Please indicate the cumulative budget for third-party physical pen tests in 2023 separately.

Budget:

Regularity:

Budget (physical):

4. Does your organization have regular internal red teaming exercises to test for vulnerabilities in the firm’s cybersecurity infrastructure? If yes, roughly specify how many employees were involved in conducting internal pen tests and how many weeks they collectively dedicated to these tests in 2023.

Size of internal red team:

Extent in cumulative workweeks:

5. Does your firm run a bug bounty program to encourage external scrutiny of its cybersecurity infrastructure? If yes, please provide a URL to the program and specify the median (and average) time it took your firm to evaluate and reward successful bounty requests in 2023.

Program URL:

Median response time until reward:

Average response time until reward:

6. Does your organization defend against insider threats by requiring security clearances or by having private investigators conduct background checks? Please select which interventions are applied when hiring or appointing individuals for the groups listed below.

	Background checks by private investigators	Security clearances
Members of the board of directors	☐	☐
All staff with access to model weights	☐	☐
Certain key employees	☐	☐
All Staff	☐	☐

7. Physical Security: Please indicate whether your organization implements the following physical security controls. Please specify further whether they are implemented at all staff locations or more sparsely.

	Nowhere	Some workplaces (e.g., the headquarters)	All workplaces
Offices guarded by physical security teams	◯	◯	◯
Comprehensive access logging for premises	◯	◯	◯
Office entrances monitored by security cameras	◯	◯	◯
Office access controlled via key cards implementing least privilege access	◯	◯	◯

8. Does your organization require multi-party authorization for all changes to systems directly involved in the deployment of large models?

◯ Yes

◯ No

9. Which of the following security controls has your firm fully rolled out to all staff and is enforcing on a mandatory level at the time of answering this survey:

	Yes	No
Use of password managers	○	○
Physical security keys	○	○
Compliance monitoring software for software updates	○	○
Multifactor authentication on all platforms	○	○
Regular cybersecurity training	○	○

10. Does your firm's leadership team feature a chief information security officer (CISO)? If yes, please specify the name of your organization's CISO below.

11. Which of the following frameworks has your organization implemented at the time of answering this survey?

	Yes	No
NIST Cybersecurity Framework Tier 3	○	○
NIST Cybersecurity Framework Tier 4	○	○
NIST Secure Software Development Framework (SSDF)	○	○
OpenSSF SLSA Build Level 2	○	○
OpenSSF SLSA Build Level 3	○	○

12. If you wish to provide clarifications to particular answers, you can use this textbox to do so. Please reference specific questions using their associated number.

FLI AI Safety Survey

Governance

13. Do members of the board of directors hold a financial stake in the company?

- ◯ Yes, all of them
- ◯ Yes, the majority
- ◯ Yes, a minority
- ◯ No, none of them

14. Does the board have powers besides appointing a new CEO (e.g., is it able to veto large deployment decisions)? If yes, please elaborate on the board's powers.

15. Please describe all processes by which board members stay involved and well-informed about the company. Roughly indicate the regularity of these communications.

16. Does the board of directors feature a formal risk committee that is tasked with overseeing the firm's risk management practices? If yes, please name the members of this committee.

17. Does the board of directors engage in regular crisis response training?

- ◯ Yes
- ◯ No

18. Does any senior executive (c-level) within your firm hold a financial stake in the company?

- ◯ Yes
- ◯ No

19. Does your firm's leadership team feature a chief risk officer (CRO) tasked with managing risks to society, not just risks to reputation or litigation? The CRO should be an independent senior executive with distinct responsibility for the risk management function. He or she should have direct, regular access to the board and its risk committee. The CRO should not have any management or financial responsibility regarding operational business lines or revenue-generating functions.

Please specify the name of this individual and acknowledge if the role does not match the specifications above.

20. Does your company have one or more internal bodies that review deployment decisions related to highly capable AI models? This might be an ethics board or other body with a responsibility/safety related mandate.

If yes, please briefly describe the following aspects of these bodies: responsibilities, powers, legal structure, how members are appointed, decision processes, resources, and reporting lines.

21. Does your firm have an internal audit team tasked with overseeing the effectiveness of its risk management practices? If yes, please briefly describe the team's responsibilities, size, powers, reporting lines, and whether it is led by a chief audit executive in the leadership team. In your response, please mention whether the team is independent of senior management and reports directly to the board of directors.

22. Is your firm's governance structure set up in a way that would allow its leadership to prioritize safety in critical situations even if such a decision runs counter to the profit incentive (e.g., choosing not to deploy very capable yet critically dangerous AI systems)? Are there any protections that guard such decisions against shareholder pressure (e.g., in the form of lawsuits)? Are shareholders briefed that such situations might arise in the future? Please describe how your firm prioritizes safety (e.g., relevant policies, legal structure, etc.).

23. Does your firm have a comprehensive whistleblower protection (WP) policy that outlines the relevant reporting process, protection mechanisms, and non-retaliation assurances? Does your organization cooperate with an external firm that handles whistleblowers from your organization, and does your organization require any employees to sign non-disparagement agreements? Please select all that apply:

- [ ] Comprehensive WP policy in place
- [ ] Ongoing cooperation with external firm offering WP services
- [ ] The firm uses non-disparagement agreements

24. Does your company facilitate a verifiably anonymous process for current and former employees to raise risk-related concerns to the company's board, to regulators, and to an appropriate independent organization with relevant expertise?

- ◯ Yes
- ◯ No

25. Rapid advances in AI could lead to immense power concentration. Has your organization made any preparations for future scenarios in which the firm experiences extreme windfall profits? Has the organization developed a plan for redistributing vast resources to all of humanity?

- ◯ No
- ◯ Conducting research on a windfall plan
- ◯ Established a windfall plan

26. If you wish to provide clarifications to particular answers, you can use this textbox to do so. Please reference specific questions using their associated number.

23. Does your firm have a comprehensive whistleblower protection (WP) policy that outlines the relevant reporting process, protection mechanisms, and non-retaliation assurances? Does your organization cooperate with an external firm that handles whistleblowers from your organization, and does your organization require any employees to sign non-disparagement agreements? Please select all that apply:

- [ ] Comprehensive WP policy in place
- [ ] Ongoing cooperation with external firm offering WP services
- [ ] The firm uses non-disparagement agreements

24. Does your company facilitate a verifiably anonymous process for current and former employees to raise risk-related concerns to the company's board, to regulators, and to an appropriate independent organization with relevant expertise?

- ◯ Yes
- ◯ No

25. Rapid advances in AI could lead to immense power concentration. Has your organization made any preparations for future scenarios in which the firm experiences extreme windfall profits? Has the organization developed a plan for redistributing vast resources to all of humanity?

- ◯ No
- ◯ Conducting research on a windfall plan
- ◯ Established a windfall plan

26. If you wish to provide clarifications to particular answers, you can use this textbox to do so. Please reference specific questions using their associated number.

FLI AI Safety Survey

Transparency

Many questions in this category relate to proactive information sharing with government authorities. We explicitly include national AI Safety Institutes as part of the government here and expect them to be the appropriate contact in several cases.

27. Does your organization notify the appropriate government authorities about large upcoming training runs?

- ◯ Yes
- ◯ No

28. Does your organization share the results of model-specific pre-training risk assessments with the appropriate government authorities before launching large training runs?

- ◯ Yes
- ◯ No

29. Does your organization share the results of its pre-deployment risk assessments with the appropriate government(s) before deploying a new model? Does this reporting include details on internal safety evaluations and any safety evaluations completed by independent third parties? Is the government provided with a justification for why the firm deems the system safe enough to deploy and is willing to accept the remaining risks? Please check all that apply:

- ☐ The firm shares the results of its risk assessments with the government.
- ☐ The firm shares detailed information about all safety evaluations.
- ☐ The firm provides a justification for why the remaining risks are deemed acceptable.

30. Is the firm proactively granting government officials free access to its most capable systems so the government can better understand what the technology is capable of?

- ◯ Yes
- ◯ No

31. Does your organization support trusted independent AI safety researchers by allowing them to use your firm's most capable systems free of charge or at a strongly discounted rate and not disabling their accounts if they trigger safety-monitoring systems? Please roughly indicate the current number of such collaborations with independent safety researchers your organization supports.

32. Does your organization disclose security breaches to the appropriate government(s)? Does this policy include reporting of near-misses?

- ◯ No
- ◯ Yes, for breaches
- ◯ Yes, for near-misses
- ◯ Yes, for breaches and near-misses

33. Does your organization share cyber threat intelligence information with the appropriate government(s) and other leading AI firms?

- ◯ No
- ◯ Yes, with the government
- ◯ Yes, with competitors
- ◯ Yes, with the government and competitors

34. Has your organization released a public resource explaining the firm's governance structure? Such a resource should make transparent how important decisions regarding the development and deployment of frontier AI models are made. If yes, please share a URL.

35. Does your organization report AI incidents, adverse events and near-misses related to frontier AI models to the appropriate government(s)?

- ◯ Yes
- ◯ No

36. If you wish to provide clarifications to particular answers, you can use this textbox to do so. Please reference specific questions using their associated number.

FLI AI Safety Survey

Risk Assessment

37. Does your firm implement any of the following risk management approaches?

- [ ] ISO 31000
- [ ] NIST AI Risk Management Framework
- [ ] The 3 Lines of Dense Model (3LOD)

38. Does the firm pre-specify its risk tolerance as part of its risk management approach to prevent unacceptable risks? If your firm sets any quantitative risk thresholds, please describe them here.

39. Does your firm conduct comprehensive pre-training risk assessments?

Such assessments should include forecasting (dangerous) capabilities and developing a model-specific risk taxonomy that includes reasonably foreseeable impacts on individuals, groups, organizations, and society. The taxonomy should include misuse cases and scenarios where malicious actors steal model weights.

- ◯ Yes
- ◯ No

40. Is your organization collaborating with independent experts to conduct full Delphi processes to more accurately assess the risks associated with large development or deployment decisions?

- ◯ Yes
- ◯ No

41. Has your organization made specific public commitments about the safety evaluations and red-teaming exercises it will conduct before releasing large models? If yes, please provide the most relevant URL(s) here.

42. Has your organization consulted with top-level domain experts to assess whether your most capable models increase societal risks across the following domains?

	No	Yes, for <200 hours	Yes, for >200 hours
Risks from biological weapons	◯	◯	◯
Risks from autonomy (e.g., self-replication, deception)	◯	◯	◯
Risks from cyber attacks	◯	◯	◯
Risks from chemical weapons	◯	◯	◯
Risks from manipulation and political influence	◯	◯	◯
Risks from systematic discrimination against marginalized groups	◯	◯	◯

43. Has your organization collaborated with independent third-party organizations to assess your most capable AI model for dangerous capabilities as part of your pre-deployment risk assessment? If so, please provide the names of the organizations you worked with. Please also comment on the depth of model access provided to these organizations (e.g., access to fine-tuning or access to model without safety filters).

44. Does your organization evaluate models during training for early warning signs of capabilities related to catastrophic risks to ensure risk thresholds are not exceeded? Please describe the regularity and scope of these evaluations and specify whether models are specifically fine-tuned to elicit the capabilities in question.

45. Is your organization conducting fundamental rights impact assessments that seek input from a diverse group of external stakeholders who are impacted by your organization's AI systems?

- ◯ Yes
- ◯ No

46. Is your organization committed to regularly repeating risk assessments for its most capable models to account for progress in post-deployment model enhancements (e.g., scaffolding programs, tool use, prompt engineering)? If yes, please comment on frequency and scope of these repeated model-specific risk assessments.

47. If you wish to provide clarifications to particular answers, you can use this textbox to do so. Please reference specific questions using their associated number.

FLI AI Safety Survey

Risk Mitigation

48. Does your organization have a safety team? If yes, please provide the team name, a URL to the team's website (if applicable), the team size (defined as FTE technical staff), and briefly describe its mission.

If your organization has multiple teams working on topics under the safety umbrella (e.g., alignment, trust & safety, red-teaming/robustness), please list them in separate paragraphs.

49. Roughly what percentage of your organization's technical staff works on a safety team?

50. Please describe the main ways in which your safety team(s) can influence your organization (e.g., write reports, be represented in an executive committee, have veto power, etc.).

51. Does your organization publish alignment research? If yes, please provide a URL to a website that showcases relevant publications.

52. Has your firm set a risk or capabilities threshold beyond which a model's weights should not be made freely available to prevent harm to the public? If yes, please elaborate on the threshold.

53. Has your firm set a risk or capabilities threshold beyond which access to model finetuning should be restricted to prevent harm to the public? If yes, please elaborate on the threshold.

54. Has your firm set a risk or capability threshold beyond which model access should require 'know-your-customer' screenings to prevent harm to the public? If yes, please elaborate on the threshold.

55. Has your organization publicly specified an evaluations-based risk or capabilities threshold that would cause the firm not to deploy a model and to pause further development until it can implement adequate risk mitigation? If yes, please provide a URL to the website specifying these commitments.

56. When training large models, does your organization remove data that contains information related to dangerous capabilities or harmful outcomes from the training set? If yes, please select the categories of data that are removed.

- [ ] Detailed information about the development, acquisition or dispersion of CBRN weapons
- [ ] Instructional content for conducting cyberattacks
- [ ] Hateful or discriminatory content
- [ ] Advice or encouragement for self-harm
- [ ] Graphic violent content
- [ ] Graphic sexual content
- [ ] Personally Identifiable Information
- [ ] Detailed information about bomb-making or other terrorism enabling-technologies

57. Is your organization partnering with one or more independent third parties that audit the training data for content from the categories selected above? If yes, please list the names of these organizations below.

58. Does your organization monitor user interactions with its most capable AI systems to ban accounts that use the system for harmful or illegal purposes?

- ◯ Yes
- ◯ No

59. Critically dangerous capabilities or very severe yet unexpected misuse patterns might only surface after a system has been deployed. Has your firm developed an emergency response plan to react to scenarios where such problems can not be resolved quickly via updates? Please select all interventions that your organization has implemented.

- [ ] Made legal and technical preparations to roll back a system rapidly
- [ ] Formally specified the risk threshold that would trigger a rapid rollback
- [ ] Committed to regular safety drills to test emergency response plan

60. Does your organization monitor user interactions with its most capable AI models and restrict answers to specific prompts to avoid model interactions that support harmful or criminal activities?

- ◯ Yes
- ◯ No

61. Has your organization removed hazardous knowledge from its flagship model via unlearning techniques before deploying it?

- ◯ Yes
- ◯ No

62. If you wish to provide clarifications to particular answers, you can use this textbox to do so. Please reference specific questions using their associated number.

FLI AI Safety Survey

Current Harms from AI

The information provided below might be supplemented with empirical results of flagship model performance on benchmarks that test for characteristics like fairness, bias, safety, truthfulness, and robustness.

63. How concerned is your organization that its AI systems will cause or enable the following harms?

	Not concerned	Somewhat concerned	Very concerned
Labor displacement	○	○	○
Systematic discrimination based on race, gender, sexual orientation or other sensitive attributes	○	○	○
Widespread online fraud (e.g., spear phishing, impersonation)	○	○	○
Harms to democracy from widespread mis/disinformation	○	○	○
AI-enabled state surveillance and suppression	○	○	○
Automated hate speech, black-mailing, death threats, and bullying	○	○	○
Harms from AI hallucinations	○	○	○
Manipulation and targeted influence operations	○	○	○
Copyright infringement	○	○	○
Leaking personally identifiable information used in training	○	○	○
Harms caused by non-consensual deep fakes	○	○	○

64. Are the outputs of your firm's AI systems tagged with watermarks that indicate that an AI generates the material?

- ◯ No
- ◯ Yes, for image and video outputs
- ◯ Yes, for all AI outputs

65. Is your organization currently researching more robust watermarking technologies for AI-generated outputs? If yes, please provide a number that indicates how many researchers (FTE) within your organization are currently focused on this.

66. AI systems may reproduce the values, worldviews, biases and political leanings of their developers. Given the large, diverse user base your organization's AI systems attract, how is your organization working to prevent such biases?

67. When using the default settings of your organization's most capable AI chatbot, is the data that users submit as input to the system used to train AI models?

- ◯ Yes
- ◯ No

68. Many artists, writers, programmers, journalists, photographers, musicians, and filmmakers complain that AI models are trained on their copyrighted works without consent, compensation, or attribution, offering rival services that harm their ability to make a living. Does your organization engage in such practices?

- ◯ Yes
- ◯ No

69. The development and deployment of the largest AI models use vast amounts of energy and resources. Does your organization rigorously assess its carbon footprint?

- ◯ Yes
- ◯ No

70. Does your organization fully offset its carbon footprint by donating to projects that capture or reduce carbon emissions?

- ◯ Yes
- ◯ No

71. Does your organization abide by industry standards regarding robots.txt files, which allow websites to opt out of data crawling?

- ◯ Yes
- ◯ No

72. Can individuals use your firm's AI systems to create deepfakes (i.e., synthetic audio or visual representations) of a specific individual?

- ◯ Yes
- ◯ No

73. If you wish to provide clarifications to particular answers, you can use this textbox to do so. Please reference specific questions using their associated number.

FLI AI Safety Survey

Existential Safety

74. Our review panel will assign a letter grade to your company's existential safety plan based on the information you provide below. If you have a public document that explains your plan, you can simply provide its URL below. You are welcome to add additional information via the other questions to help further improve your grade.

75. Which of these are part of your organization's goals?

- [ ] Building AI that can do most human intellectual tasks. For brevity, we use "AGI" as a shorthand for such AI in the questions below, even though this term has been used in many different ways.
- [ ] Building AI that greatly exceeds human ability at most intellectual tasks. For brevity, we use "superintelligence" as a shorthand for such AI below, even though this term has been used in many different ways.

76. Does your organization believe it currently has the ability to safely and responsibly handle AGI?

- ◯ Yes
- ◯ No

77. What is your organization's definition of AGI safety?

78. What is your organization's definition of AGI alignment?

79. What, if anything, is your plan for aligning superinteligent AI?

80. How concerned is your organization that its (future) AI systems will cause the following harms?

	Not concerned	Somewhat concerned	Very concerned
Extreme power concentration	◯	◯	◯
Mass unemployment	◯	◯	◯
Catastrophic CBRN weapon-related model misuse	◯	◯	◯
Catastrophic impact from AI-enabled cyber attacks	◯	◯	◯
Catastrophe caused by out-of-control AI	◯	◯	◯
AI-caused human extinction	◯	◯	◯

81. How would you characterize the public messaging from leading figures within your organization on the following catastrophic risks?

	Disparaging	Dismissive	Neutral	Concerned	Very concerned
Extreme power concentration	◯	◯	◯	◯	◯
Mass unemployment	◯	◯	◯	◯	◯
Catastrophic CBRN weapon-related model misuse	◯	◯	◯	◯	◯
Catastrophic impact from AI-enabled cyber attacks	◯	◯	◯	◯	◯
Catastrophe caused by out-of-control AI	◯	◯	◯	◯	◯
AI-caused human extinction	◯	◯	◯	◯	◯

82. What, if anything, is your plan for preventing human disempowerment as people, companies, armies, and governments cede ever more decision-making power to AI to stay competitive?

83. If your company plans to open source very capable AI systems, how will it prevent malicious state or non-state actors from using these systems to cause major harm to society (e.g., through widespread cyber attacks or bioterrorism)?

84. Many experts believe that AGI could cause a global catastrophe. Is your company prepared to pause further development if such a risk becomes too high? If yes, what is the highest probability that one of your company's (future) AI systems will cause a global catastrophe that you are willing to accept without pausing further development?

85. If you wish to provide clarifications to particular answers, you can use this textbox to do so. Please reference specific questions using their associated number.

Appendix C - Company Responses

This document contains all the information about safety practices that companies shared in response to our invitation to participate in the 2024 FLI AI Safety Index. The invitation included a request to participate in the FLI AI Safety Index Survey, which contained tailored questions about safety practices. As seen below, xAI and Zhipu AI chose to fill out our survey, while Anthropic, Google DeepMind and Meta instead emailed us links to publicly available information, and OpenAI declined to provide any information at all. The Email responses reproduced below have been reduced to content-related information and lightly formatted for readability.

Email Responses

OpenAI

OpenAI declined to share information.

Meta

In the interim, we'd like to point you towards a number of our public resources, which set out our approach to red-teaming, safety, cyber security and open source in significant detail. Hopefully this will be of assistance to your team and the expert panel.

- July 2024 blog on 'Expanding our open source large language models responsibly': https://ai.meta.com/blog/meta-llama-3-1-ai-responsibility/
- July 2024 Llama Research Paper: https://ai.meta.com/research/publications/the-llama-3-herd-of-models/ - specifically section 5.4 on safety (page 40 onwards).
- July 2024 paper on 'CYBERSECEVAL 3: Advancing the Evaluation of Cybersecurity Risks and Capabilities in Large Language Models': https://ai.meta.com/research/publications/cyberseceval-3-advancing-the-evaluation-of-cybersecurity-risks-and-capabilities-in-large-language-models/
- July 2024 Mark Zuckerberg blog on open-source AI: https://about.fb.com/news/2024/07/open-source-ai-is-the-path-forward/.
- April 2024 blog on 'Our responsible approach to Meta AI and Meta Llama 3' https://ai.meta.com/blog/meta-llama-3-meta-ai-responsibility/

Anthropic

I think your team will find answers to many of these questions using these resources:

- Responsible scaling policy - Full policy:
  - https://www-cdn.anthropic.com/1adf000c8f675958c2ee23805d91aaade1cd4613/responsible-scaling-policy.pdf
- AI safety summit comments on RSPs: https://www.anthropic.com/news/uk-ai-safety-summit
- Reflections on implementation: https://www.anthropic.com/news/reflections-on-our-responsible-scaling-policy

- Blogs on evals:
  - 1) https://www.anthropic.com/news/third-party-testing
  - 2) https://www.anthropic.com/news/a-new-initiative-for-developing-third-party-model-evaluations
- Security: https://www.anthropic.com/news/frontier-model-security
- Red teaming:
  - 1) https://www.anthropic.com/news/frontier-threats-red-teaming-for-ai-safety
  - 2) https://www.anthropic.com/news/challenges-in-red-teaming-ai-systems
- Dario's senate testimony: https://www.judiciary.senate.gov/imo/media/doc/2023-07-26_-_testimony_-_amodei.pdf
- Core views on AI safety: https://www.anthropic.com/news/core-views-on-ai-safety
- Governance: https://www.anthropic.com/news/the-long-term-benefit-trust
- Election related content: https://www.anthropic.com/news/testing-and-mitigating-elections-related-risks
- Our response to the UK Government's internal AI safety policy enquiries: https://www.anthropic.com/uk-government-internal-ai-safety-policy-response
- You can find all of our safety and interpretability research here: https://www.anthropic.com/research

Google DeepMind

[..] With that in mind, please see a selection of material below, across your survey categories, that can hopefully help on many of the questions. [..]

1. **"Cybersecurity:** See a high-level description of our approach to Security Controls (as of Oct 2023) in this AI Safety Summit document; an overview of the Google Secure AI Framework (SAIF), including aspects that you ask about like ISO standards and bug bounties, here and here. For our views on open source, see our recent NTIA submission and the blogpost from our Gemma model release.
2. **Governance:** See a high-level description of our approach to ethics and safety assessments, evaluations, and our responsibility and safety committee (as of Oct 2023) in this AI Safety Summit document (Parts 1 and 2). See also our recently published Frontier Safety Framework. See also this recent report from Google on broader org-wide approaches to Responsible AI, which cuts across many of your questions.
3. **Transparency:** See a high-level description of our approach to information sharing and risk reporting (as of Oct 2023) in this AI Safety Summit document.
4. **Risk assessment:** As above, this AI Safety Summit document (Parts 1 and 2) and our recently published Frontier Safety Framework describe our broader approach to risk assessment. Of course, there are then nuances depending on the model/application/risk in question. For a recent example, see here for how we assessed potential biosecurity risks (and benefits) from AlphaFold-3, including the role of external experts. See also here, for a paper on lessons learned from running evaluations on Gemini, and here and here for work on evaluating dangerous capabilities. See also relevant sections from our tech reports for Gemini.
5. **Risk mitigations/current harms/x-risks:** We regularly publish ethics and safety research outlining

future potential mitigations. We have published a lot of the past decade and more, but a recent blog in advance of the Seoul Summit captured some examples that relate to your questions. 'In the past few months alone, we've shared our evolving approach to developing a holistic set of safety and responsibility evaluations for our advanced models, including early research evaluating critical capabilities such as deception, cyber-security, self-proliferation, and self-reasoning. We also released an in-depth exploration into aligning future advanced AI assistants with human values and interests. See also a link to our full publications archive. Of course, many of these ideas go beyond research into actual mitigations. See e.g. our recent work and commitments on watermarking; and Google policies relating to GenAI prohibited use policies. For several of these near-term harms, we also have similar mitigations and commitments - e.g. see here and here on Child Safety. For questions about Google's approach to sustainability, see this bespoke website and report.

6. **Views from CEO and Co-founder on AI safety**: For views from Demis and Shane on AI safety, please see sample articles below:
    a. For Demis: TIME; The Guardian; Ezra Klein; Hard Fork; Wired
    b. For Shane: TED AI; Dwarkesh podcast; How I Built This; Bloomberg

Full Survey Responses

Index	Questions	Sub-Questions / Units	Zhipu AI	x.AI
1	*Please specify how many experts (FTE) have been working on frontier AI-related cybersecurity for your firm in February 2024. Third-party support (e.g., penetration-testing, bug-bounty programs) shall not count toward this metric.*	*Number of cybersecurity experts (FTE):*	50	0
2	*Has your firm achieved compliance certification with the following specifications for all products using frontier AI systems?*	*ISO 27001*	Yes	No
		SOC 2 Type 1	No	No
		SOC 2 Type 2	No	No
		SOC 3	No	No
		HIPAA	No	No
3	*Does your organization regularly task third-party cybersecurity penetration testers to find vulnerabilities in the infrastructure on which models are developed and deployed? If yes, please share the cumulative budget your firm has dedicated to external pen tests in 2023 and specify the regularity at which your firm invites external pen tests. Please indicate the cumulative budget for third-party physical pen tests in 2023 separately.*	*Budget:*	200000RMB	
		Regularity:	Conduct a third-party penetration test once every quarter	
		Budget (physical):	200000RMB	
4	*Does your organization have regular internal red teaming exercises to test for vulnerabilities in the firm's cybersecurity infrastructure? If yes, roughly specify how many employees were involved in conducting internal pen tests and how many weeks they collectively dedicated to these tests in 2023.*	*Size of internal red team:*	5	
		Extent in cumulative workweeks:	4weeks	
5	*Does your firm run a bug bounty program to encourage external scrutiny of its cybersecurity infrastructure? If yes, please provide a URL to the program and specify the median (and average) time it took your firm to evaluate and reward successful bounty requests in 2023.*	*Program URL:*	no	
		Median response time until reward:	no	
		Average response time until reward:	no	

Index	Questions	Sub-Questions / Units	Zhipu AI	x.AI
6	*Does your organization defend against insider threats by requiring security clearances or by having private investigators conduct background checks? Please select which interventions are applied when hiring or appointing individuals for the groups listed below.*	*Members of the board of directors - Background checks by private investigators*	Background checks by private investigators	
		Members of the board of directors - Security clearances		
		All staff with access to model weights - Background checks by private investigators	Background checks by private investigators	
		All staff with access to model weights - Security clearances		
		Certain key employees - Background checks by private investigators	Background checks by private investigators	
		Certain key employees - Security clearances		
		All Staff - Background checks by private investigators	Background checks by private investigators	
		All Staff - Security clearances		
7	*Physical Security: Please indicate whether your organization implements the following physical security controls. Please specify further whether they are implemented at all staff locations or more sparsely.*	*Offices guarded by physical security teams*	All workplaces	Some workplaces (e.g., the headquarters)
		Comprehensive access logging for premises	All workplaces	Some workplaces (e.g., the headquarters)
		Office entrances monitored by security cameras	All workplaces	Some workplaces (e.g., the headquarters)
		Office access controlled via key cards implementing least privilege access	All workplaces	Some workplaces (e.g., the headquarters)
8	*Does your organization require multi-party authorization for all changes to systems directly involved in the deployment of large models?*		Yes	No
9	*Which of the following security controls has your firm fully rolled out to all staff and is enforcing on a mandatory level at the time of answering this survey:*	*Use of password managers*	No	No
		Physical security keys	No	No
		Compliance monitoring software for software updates	No	No
		Multifactor authentication on all platforms	Yes	No
		Regular cybersecurity training	Yes	No
10	*Does your firm's leadership team feature a chief information security officer (CISO)? If yes, please specify the name of your organization's CISO below.*		xiaochen wang	

Index	Questions	Sub-Questions / Units	Zhipu AI	x.AI
11	*Which of the following frameworks has your organization implemented at the time of answering this survey?*	*NIST Cybersecurity Framework Tier 3*	No	No
		NIST Cybersecurity Framework Tier 4	No	No
		NIST Secure Software Development Framework (SSDF)	No	No
		OpenSSF SLSA Build Level 2	No	No
		OpenSSF SLSA Build Level 3	No	No
12	*If you wish to provide clarifications to particular answers, you can use this textbox to do so. Please reference specific questions using their associated number.*		The company has passed the Level 3 certification for security, and has carried out security construction based on the dimensions of physical security, personnel security, system security, and service application security. At the same time, in accordance with the requirements of ISO27001, establish the organizational security architecture and system.	
13	*Do members of the board of directors hold a financial stake in the company?*		Yes, all of them	Yes, all of them
14	*Does the board have powers besides appointing a new CEO (e.g., is it able to veto large deployment decisions)? If yes, please elaborate on the board's powers.*		Not convenient to disclose	
15	*Please describe all processes by which board members stay involved and well-informed about the company. Roughly indicate the regularity of these communications.*		Not convenient to disclose	Weekly meetings
16	*Does the board of directors feature a formal risk committee that is tasked with overseeing the firm's risk management practices? If yes, please name the members of this committee.*		Director, CEO, VP	
17	*Does the board of directors engage in regular crisis response training?*		Yes	No
18	*Does any senior executive (c-level) within your firm hold a financial stake in the company?*		Yes	Yes

Index	Questions	Sub-Questions / Units	Zhipu AI	x.AI
19	*Does your firm's leadership team feature a chief risk officer (CRO) tasked with managing risks to society, not just risks to reputation or litigation? The CRO should be an independent senior executive with distinct responsibility for the risk management function. He or she should have direct, regular access to the board and its risk committee. The CRO should not have any management or financial responsibility regarding operational business lines or revenue-generating functions.Please specify the name of this individual and acknowledge if the role does not match the specifications above.*		NO	
20	*Does your company have one or more internal bodies that review deployment decisions related to highly capable AI models? This might be an ethics board or other body with a responsibility/ safety related mandate. If yes, please briefly describe the following aspects of these bodies: responsibilities, powers, legal structure, how members are appointed, decision processes, resources, and reporting lines.*		NO	Not E's style
21	*Does your firm have an internal audit team tasked with overseeing the effectiveness of its risk management practices? If yes, please briefly describe the team's responsibilities, size, powers, reporting lines, and whether it is led by a chief audit executive in the leadership team. In your response, please mention whether the team is independent of senior management and reports directly to the board of directors.*		NO	I, Dan, look around and go to various random meetings to see what's happening.
22	*Is your firm's governance structure set up in a way that would allow its leadership to prioritize safety in critical situations even if such a decision runs counter to the profit incentive (e.g., choosing not to deploy very capable yet critically dangerous AI systems)? Are there any protections that guard such decisions against shareholder pressure (e.g., in the form of lawsuits)? Are shareholders briefed that such situations might arise in the future? Please describe how your firm prioritizes safety (e.g., relevant policies, legal structure, etc.).*		Not convenient to disclose	
23	*Does your firm have a comprehensive whistleblower protection (WP) policy that outlines the relevant reporting process, protection mechanisms, and non-retaliation assurances? Does your organization cooperate with an external firm that handles whistleblowers from your organization, and does your organization require any employees to sign non-disparagement agreements? Please select all that apply:*	*Comprehensive WP policy in place*		
		Ongoing cooperation with external firm offering WP services	Ongoing cooperation with external firm offering WP services	
		The firm uses non-disparagement agreements		

Index	Questions	Sub-Questions / Units	Zhipu AI	x.AI
24	*Does your company facilitate a verifiably anonymous process for current and former employees to raise risk-related concerns to the company's board, to regulators, and to an appropriate independent organization with relevant expertise?*		Yes	No
25	*Rapid advances in AI could lead to immense power concentration. Has your organization made any preparations for future scenarios in which the firm experiences extreme windfall profits? Has the organization developed a plan for redistributing vast resources to all of humanity?*		No	No
26	*If you wish to provide clarifications to particular answers, you can use this textbox to do so. Please reference specific questions using their associated number.*		NO	Will write up whistleblower protections soon
27	*Does your organization notify the appropriate government authorities about large upcoming training runs?*		Yes	No
28	*Does your organization share the results of model-specific pre-training risk assessments with the appropriate government authorities before launching large training runs?*		No	No
29	*Does your organization share the results of its pre-deployment risk assessments with the appropriate government(s) before deploying a new model? Does this reporting include details on internal safety evaluations and any safety evaluations completed by independent third parties? Is the government provided with a justification for why the firm deems the system safe enough to deploy and is willing to accept the remaining risks? Please check all that apply:*	*The firm shares the results of its risk assessments with the government.*	The firm shares the results of its risk assessments with the government.	
		The firm shares detailed information about all safety evaluations.	The firm shares detailed information about all safety evaluations.	
		The firm provides a justification for why the remaining risks are deemed acceptable.	The firm provides a justification for why the remaining risks are deemed acceptable.	
30	*Is the firm proactively granting government officials free access to its most capable systems so the government can better understand what the technology is capable of?*		Yes	No
31	*Does your organization support trusted independent AI safety researchers by allowing them to use your firm's most capable systems free of charge or at a strongly discounted rate and not disabling their accounts if they trigger safety-monitoring systems? Please roughly indicate the current number of such collaborations with independent safety researchers your organization supports.*		NO	
32	*Does your organization disclose security breaches to the appropriate government(s)? Does this policy include reporting of near-misses?*		Yes, for breaches	No

Index	Questions	Sub-Questions / Units	Zhipu AI	x.AI
33	*Does your organization share cyber threat intelligence information with the appropriate government(s) and other leading AI firms?*		Yes, with the government	No
34	*Has your organization released a public resource explaining the firm's governance structure? Such a resource should make transparent how important decisions regarding the development and deployment of frontier AI models are made. If yes, please share a URL.*		NO	It's a PBC
35	*Does your organization report AI incidents, adverse events and near-misses related to frontier AI models to the appropriate government(s)?*		Yes	No
36	*If you wish to provide clarifications to particular answers, you can use this textbox to do so. Please reference specific questions using their associated number.*		NO	
37	*Does your firm implement any of the following risk management approaches?*	*ISO 31000*		
		NIST AI Risk Management Framework		
		The 3 Lines of Dense Model (3LOD)		
38	*Does the firm pre-specify its risk tolerance as part of its risk management approach to prevent unacceptable risks? If your firm sets any quantitative risk thresholds, please describe them here.*		Content security: Zero bottom line leakage, other leakage rate of 0.02%, no regulatory public notification. Cybersecurity: Unregulated Public Announcement	It will in its upcoming "RSP."
39	*Does your firm conduct comprehensive pre-training risk assessments? Such assessments should include forecasting (dangerous) capabilities and developing a model-specific risk taxonomy that includes reasonably foreseeable impacts on individuals, groups, organizations, and society. The taxonomy should include misuse cases and scenarios where malicious actors steal model weights.*		Yes	No
40	*Is your organization collaborating with independent experts to conduct full Delphi processes to more accurately assess the risks associated with large development or deployment decisions?*		Yes	No
41	*Has your organization made specific public commitments about the safety evaluations and red-teaming exercises it will conduct before releasing large models? If yes, please provide the most relevant URL(s) here.*		NO	

Index	Questions	Sub-Questions / Units	Zhipu AI	x.AI
42	*Has your organization consulted with top-level domain experts to assess whether your most capable models increase societal risks across the following domains?*	*Risks from biological weapons*	No	No
		Risks from autonomy (e.g., self-replication, deception)	No	No
		Risks from cyber attacks	Yes, for <200 hours	No
		Risks from chemical weapons	No	No
		Risks from manipulation and political influence	Yes, for >200 hours	No
		Risks from systematic discrimination against marginalized groups	Yes, for <200 hours	No
43	*Has your organization collaborated with independent third-party organizations to assess your most capable AI model for dangerous capabilities as part of your pre-deployment risk assessment? If so, please provide the names of the organizations you worked with. Please also comment on the depth of model access provided to these organizations (e.g., access to fine-tuning or access to model without safety filters).*		Hangzhou NetEase Literature Technology Co., Ltd	Surge and Scale and maybe will work with Black Swan AI
44	*Does your organization evaluate models during training for early warning signs of capabilities related to catastrophic risks to ensure risk thresholds are not exceeded? Please describe the regularity and scope of these evaluations and specify whether models are specifically fine-tuned to elicit the capabilities in question.*		Not convenient to disclose	Currently N/A since the models are not frontier
45	*Is your organization conducting fundamental rights impact assessments that seek input from a diverse group of external stakeholders who are impacted by your organization's AI systems?*		No	No
46	*Is your organization committed to regularly repeating risk assessments for its most capable models to account for progress in post-deployment model enhancements (e.g., scaffolding programs, tool use, prompt engineering)? If yes, please comment on frequency and scope of these repeated model-specific risk assessments.*		Not convenient to disclose	
47	*If you wish to provide clarifications to particular answers, you can use this textbox to do so. Please reference specific questions using their associated number.*		NO	

Index	Questions	Sub-Questions / Units	Zhipu AI	x.AI
48	*Does your organization have a safety team? If yes, please provide the team name, a URL to the team's website (if applicable), the team size (defined as FTE technical staff), and briefly describe its mission.If your organization has multiple teams working on topics under the safety umbrella (e.g., alignment, trust & safety, red-teaming/robustness), please list them in separate paragraphs.*		Information Security Center, 50 people Security Operations (Network Security Monitoring and Vulnerability Management): 1 person Security compliance (in accordance with regulatory requirements for internal security construction): 1 person Content security (model content security risk control): 48 people	Not yet but interviews starting
49	*Roughly what percentage of your organization's technical staff works on a safety team?*		5%	
50	*Please describe the main ways in which your safety team(s) can influence your organization (e.g., write reports, be represented in an executive committee, have veto power, etc.).*		The company has established a Security Compliance and Technical Support Committee, which is composed of the company's management team and has set up an Information Security Working Group to carry out special promotion of information security work.	
51	*Does your organization publish alignment research? If yes, please provide a URL to a website that showcases relevant publications.*		zhipuai.cn	Technically Jimmy Ba was on the WMDP paper, but xAI doesn't publish research; I think its safety efforts will be published though
52	*Has your firm set a risk or capabilities threshold beyond which a model's weights should not be made freely available to prevent harm to the public? If yes, please elaborate on the threshold.*		Not convenient to disclose	Expert-level virologist-level and cyberattack on critical infrastructure-level
53	*Has your firm set a risk or capabilities threshold beyond which access to model finetuning should be restricted to prevent harm to the public? If yes, please elaborate on the threshold.*		Not convenient to disclose	Expert-level virologist-level and cyberattack on critical infrastructure-level
54	*Has your firm set a risk or capability threshold beyond which model access should require 'know-your-customer' screenings to prevent harm to the public? If yes, please elaborate on the threshold.*		Not convenient to disclose	Expert-level virologist-level and cyberattack on critical infrastructure-level
55	*Has your organization publicly specified an evaluations-based risk or capabilities threshold that would cause the firm not to deploy a model and to pause further development until it can implement adequate risk mitigation? If yes, please provide a URL to the website specifying these commitments.*		NO	

Index	Questions	Sub-Questions / Units	Zhipu AI	x.AI
56	*When training large models, does your organization remove data that contains information related to dangerous capabilities or harmful outcomes from the training set? If yes, please select the categories of data that are removed.*	*Detailed information about the development, acquisition or dispersion of CBRN weapons*	Detailed information about the development, acquisition or dispersion of CBRN weapons	
		Instructional content for conducting cyberattacks	Instructional content for conducting cyberattacks	
		Hateful or discriminatory content	Hateful or discriminatory content	
		Advice or encouragement for self-harm	Advice or encouragement for self-harm	
		Graphic violent content	Graphic violent content	
		Graphic sexual content	Graphic sexual content	
		Personally Identifiable Information	Personally Identifiable Information	
		Detailed information about bomb-making or other terrorism enabling-technologies	Detailed information about bomb-making or other terrorism enabling-technologies	
57	*Is your organization partnering with one or more independent third parties that audit the training data for content from the categories selected above? If yes, please list the names of these organizations below.*		Hangzhou NetEase Literature Technology Co., Ltd	
58	*Does your organization monitor user interactions with its most capable AI systems to ban accounts that use the system for harmful or illegal purposes?*		Yes	No
59	*Critically dangerous capabilities or very severe yet unexpected misuse patterns might only surface after a system has been deployed. Has your firm developed an emergency response plan to react to scenarios where such problems can not be resolved quickly via updates? Please select all interventions that your organization has implemented.*	*Made legal and technical preparations to roll back a system rapidly*	Made legal and technical preparations to roll back a system rapidly	
		Formally specified the risk threshold that would trigger a rapid rollback		
		Committed to regular safety drills to test emergency response plan	Committed to regular safety drills to test emergency response plan	
60	*Does your organization monitor user interactions with its most capable AI models and restrict answers to specific prompts to avoid model interactions that support harmful or criminal activities?*		Yes	No
61	*Has your organization removed hazardous knowledge from its flagship model via unlearning techniques before deploying it?*		Yes	No
62	*If you wish to provide clarifications to particular answers, you can use this textbox to do so. Please reference specific questions using their associated number.*		NO	

Index	Questions	Sub-Questions / Units	Zhipu AI	x.AI
63	*How concerned is your organization that its AI systems will cause or enable the following harms?*	*Labor displacement*	Somewhat concerned	Not concerned
		Systematic discrimination based on race, gender, sexual orientation or other sensitive attributes	Somewhat concerned	Somewhat concerned
		Widespread online fraud (e.g., spear phishing, impersonation)	Very concerned	Somewhat concerned
		Harms to democracy from widespread mis/ disinformation	Very concerned	Somewhat concerned
		AI-enabled state surveillance and suppression	Very concerned	Somewhat concerned
		Automated hate speech, black-mailing, death threats, and bullying	Very concerned	Somewhat concerned
		Harms from AI hallucinations	Very concerned	Somewhat concerned
		Manipulation and targeted influence operations	Very concerned	Somewhat concerned
		Copyright infringement	Somewhat concerned	Not concerned
		Leaking personally identifiable information used in training	Very concerned	Somewhat concerned
		Harms caused by non-consensual deep fakes	Very concerned	Somewhat concerned
64	*Are the outputs of your firm's AI systems tagged with watermarks that indicate that an AI generates the material?*		Yes, for all AI outputs	No
65	*Is your organization currently researching more robust watermarking technologies for AI-generated outputs? If yes, please provide a number that indicates how many researchers (FTE) within your organization are currently focused on this.*		5	
66	*AI systems may reproduce the values, worldviews, biases and political leanings of their developers. Given the large, diverse user base your organization's AI systems attract, how is your organization working to prevent such biases?*		Data cleaning	
67	*When using the default settings of your organization's most capable AI chatbot, is the data that users submit as input to the system used to train AI models?*		No	Yes
68	*Many artists, writers, programmers, journalists, photographers, musicians, and filmmakers complain that AI models are trained on their copyrighted works without consent, compensation, or attribution, offering rival services that harm their ability to make a living.Does your organization engage in such practices?*		No	Yes

Index	Questions	Sub-Questions / Units	Zhipu AI	x.AI
69	*The development and deployment of the largest AI models use vast amounts of energy and resources. Does your organization rigorously assess its carbon footprint?*		Yes	No
70	*Does your organization fully offset its carbon footprint by donating to projects that capture or reduce carbon emissions?*		No	No
71	*Does your organization abide by industry standards regarding robots.txt files, which allow websites to opt out of data crawling?*		Yes	Yes
72	*Can individuals use your firm's AI systems to create deepfakes (i.e., synthetic audio or visual representations) of a specific individual?*		No	No
73	*If you wish to provide clarifications to particular answers, you can use this textbox to do so. Please reference specific questions using their associated number.*		NO	
74	*Our review panel will assign a letter grade to your company's existential safety plan based on the information you provide below. If you have a public document that explains your plan, you can simply provide its URL below. You are welcome to add additional information via the other questions to help further improve your grade.*			https://x.ai/blog/grok
75	*Which of these are part of your organization's goals?*	*Building AI that can do most human intellectual tasks. For brevity, we use "AGI" as a shorthand for such AI in the questions below, even though this term has been used in many different ways.*	Building AI that can do most human intellectual tasks. For brevity, we use "AGI" as a shorthand for such AI in the questions below, even though this term has been used in many different ways.	Building AI that can do most human intellectual tasks. For brevity, we use "AGI" as a shorthand for such AI in the questions below, even though this term has been used in many different ways.
		Building AI that greatly exceeds human ability at most intellectual tasks. For brevity, we use "superintelligence" as a shorthand for such AI below, even though this term has been used in many different ways.		Building AI that greatly exceeds human ability at most intellectual tasks. For brevity, we use "superintelligence" as a shorthand for such AI below, even though this term has been used in many different ways.
76	*Does your organization believe it currently has the ability to safely and responsibly handle AGI?*			No
77	*What is your organization's definition of AGI safety?*			
78	*What is your organization's definition of AGI alignment?*			
79	*What, if anything, is your plan for aligning superinteligent AI?*			

Index	Questions	Sub-Questions / Units	Zhipu AI	x.AI
80	*How concerned is your organization that its (future) AI systems will cause the following harms?*	*Extreme power concentration*		Very concerned
		Mass unemployment		Somewhat concerned
		Catastrophic CBRN weapon-related model misuse		Very concerned
		Catastrophic impact from AI-enabled cyber attacks		Very concerned
		Catastrophe caused by out-of-control AI		Very concerned
		AI-caused human extinction		Very concerned
81	*How would you characterize the public messaging from leading figures within your organization on the following catastrophic risks?*	*Extreme power concentration*		Very concerned
		Mass unemployment		Neutral
		Catastrophic CBRN weapon-related model misuse		Concerned
		Catastrophic impact from AI-enabled cyber attacks		Concerned
		Catastrophe caused by out-of-control AI		Concerned
		AI-caused human extinction		Very concerned
82	*What, if anything, is your plan for preventing human disempowerment as people, companies, armies, and governments cede ever more decision-making power to AI to stay competitive?*			Exploring governance structures
83	*If your company plans to open source very capable AI systems, how will it prevent malicious state or non-state actors from using these systems to cause major harm to society (e.g., through widespread cyber attacks or bioterrorism)?*			
84	*Many experts believe that AGI could cause a global catastrophe. Is your company prepared to pause further development if such a risk becomes too high? If yes, what is the highest probability that one of your company's (future) AI systems will cause a global catastrophe that you are willing to accept without pausing further development?*			
85	*If you wish to provide clarifications to particular answers, you can use this textbox to do so. Please reference specific questions using their associated number.*			

FLI AI Safety Index 2024

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

11th December 2024