Future of Life Institute

AI Safety Index

Summer 2026

July 2026

Available online at: futureoflife.org/index

Contact us: policy@futureoflife.org

future of life INSTITUTE

Contents

About the Organization: The Future of Life Institute (FLI) is an independent nonprofit organization with the goal of reducing large-scale risks and steering transformative technologies to benefit humanity, with a particular focus on artificial intelligence (AI). Learn more at futureoflife.org.

1 Executive Summary

	Anthropic	OpenAI	Google DeepMind	Meta	Z.ai	Alibaba Cloud	xAI	DeepSeek	Mistral
Overall Grade	C+	C	C	D+	D-	D-	F	F	F
Score	2.66	2.28	2.01	1.32	0.88	0.87	0.65	0.47	0.33
Grade Trend (Winter 25)	C+	C+ ▼	C	D ▲	D ▼	D-	D ▼	D ▼	N/A
Risk Assessment (6 indicators)	C+	C+	C+	D+	F	F	D-	F	F
Current Harms (9 indicators)	B-	C	C	D-	C-	C-	F	D-	F
Safety Frameworks (4 indicators)	B-	C+	C	C-	D-	D-	D	F	F
Existential Safety (4 indicators)	D+	D+	D	F	F	F	F	F	F
Governance & Accountability (4 indicators)	B	C	C-	D+	F	D-	F	F	F
Information Sharing (10 indicators)	B+	B-	B-	D+	D	D	D	D-	D-
Survey Responses	✓	✓	✓	✓	✓	✗	✗	✗	✗

Grading: Uses the US GPA system for grade boundaries: A+, A, A-, B+, […], F letter values corresponding to numerical values 4.3, 4.0, 3.7, 3.3, […], 0.

1.1 Key Findings

- **Anthropic, OpenAI, and Google DeepMind stay on top.** Anthropic again earns the highest overall grade and leads five of six domains via relatively strong transparency, a comparatively established safety framework, technical research, and governance. OpenAI now leads in Risk Assessment on the strength of a broader evaluation suite and diverse engagement with external testing.
- **Meta improves and xAI deteriorates:** Meta improved from 6th to 4th place, while xAI dropped from 4th to 7th place.
- **European dissonance:** Although the European Union is a leader in AI safety regulation, the top European AI company Mistral scored dead last on safety.
- **Inadequate safety is a global problem, not a regional one.** Three companies receive failing grades, one each from the US (xAI), China (DeepSeek), and Europe (Mistral).
- **Reviewers flagged the industry's pivot to military AI use as an emerging current harm risk.** From 2024 to 2026, companies including Anthropic, OpenAI, Google DeepMind, and Meta that previously banned military applications gradually reversed course, joining xAI and Mistral in actively seeking defense partnerships. Despite their limits on domestic surveillance and autonomous weapons, Anthropic drew criticism from the review panel for "questionable military engagements," including a reported link to the Minab school strike that caused mass civilian deaths. Leading Chinese firms, meanwhile, face U.S. allegations of military ties that Alibaba Cloud and Z.ai deny.

- **Even industry leaders in safety practices are retreating from prior commitments.** Anthropic, OpenAI, Google DeepMind, and Meta have weakened or voided pledges to pause unilaterally if redlines are approached, some citing competitor-contingent conditions. Reviewers call this "moving goalpost" and argue that it has "undermined safety frameworks across the board".
- **Existential Safety is the weakest domain industry-wide.** No company exceeds C-; most score D or below. Constructive attempts exist, such as Anthropic's constitutional classifiers, OpenAI's call for governance institutions, Google DeepMind's monitoring commitments, and Meta's loss-of-control provisions, but are judged by panelists to be "entirely inadequate." Dominant paradigms such as interpretability and Chain-of-Thought (CoT) monitorability are questioned because "detection is not prevention."
- **Safety rhetoric outpaces revealed behavior.** Across Google DeepMind, OpenAI, and xAI, leadership's reassuring public messaging diverges from commercial conduct and legislative stance, making stated commitments an unreliable proxy for actual safety practice.
- **Companies are publishing and updating safety frameworks, but these frameworks have weak teeth.** As US/EU compliance deadlines near, Anthropic, OpenAI, Google DeepMind, Meta, and xAI published and updated fuller frameworks — yet they sometimes lack quantitative thresholds, genuinely independent audits, and clear decision authority.

1.2 Company Progress Highlights and Key Recommendations

Company	*Progress Highlights*	*Key Recommendations*
Anthropic	▪ Comparatively detailed safety framework with commitments for third-party audits. ▪ Solid evaluations of autonomous R&D and scheming/misalignment capabilities of frontier models with strong elicitation. ▪ Continuous industry-leading transparency with both published model specs and system prompts.	▪ Reverse the RSP 3.0 walk-back on pause commitments and restore credibility of commitments. ▪ Replace qualitative thresholds with quantitative and risk-tied ones. ▪ Treat prevention as seriously as interpretability/detection. ▪ Establish stronger safeguards for military use of its AI systems.
OpenAI	▪ Strong external testing and comparatively broad risk assessment. ▪ Called for global governance institutions to slow development when needed. ▪ Regular reports documenting their disruption of malicious uses of their AI systems.	▪ Remove leadership's ability to override the Safety Advisory Group. ▪ Make safety-framework thresholds measurable, risk-tiered, and externally enforceable, with commitment to notify authorities when risk thresholds are crossed. ▪ Evaluate internal-deployment risks before broad internal use rather than after. ▪ Align public-policy positions with stated safety commitments.
Google DeepMind	▪ Updated Frontier Safety Framework adding manipulation, misalignment, and internal-deployment coverage. ▪ Strong watermark protection.	▪ Establish clear decision-making authority, an executive risk officer, and independent audit. It remains unclear which internal body can halt deployment independently of executive leadership. ▪ Make safety-framework thresholds measurable and risk-tiered. ▪ Reverse the backsliding on pause commitments. ▪ Align public-policy positions with stated safety commitments from leadership.

Company	Progress Highlights	Key Recommendations
Meta	▪ Published safety framework with more details of risk identification and threat modeling. ▪ Bug bounties cover catastrophic risk factors.	▪ Strengthen whistleblower protections and align culture with policy. The whistleblowing policy quality scores are reasonable but undermined by active enforcement of a non-disparagement agreement and other suppression of dissent. ▪ Make safety-framework thresholds measurable and risk-tiered. ▪ Establish auditing mechanisms for the safety framework.
Z.ai	▪ More transparent than its Chinese peers, with some proactive safety research built into products. ▪ Some meaningful incident-response infrastructure and internal-deployment threat mitigation.	▪ Publish a full safety framework and governance structure. The company has no published framework; its rating largely reflects the Chinese regulatory environment rather than independent safety leadership. ▪ Move beyond passive deference to regulation toward proactive safety research. Deferring entirely to government guidance amounts to "complete passivity" as an existential-safety strategy for highly advanced AI systems. ▪ Establish and publicize a whistleblower policy. There is no clear governance structure or whistleblowing channel, even absent reported incidents.
Alibaba Cloud	▪ Stronger-than-expected benchmark performance with more transparent disclosure of misalignment propensity than its peers. ▪ Two-layered safety strategy spanning the Qwen model-level team and the Alibaba security team.	▪ Publish a full safety framework and governance structure. The company has no published framework; its rating largely reflects the Chinese regulatory environment rather than independent safety leadership. ▪ Move beyond passive deference to regulation toward proactive safety research. Deferring entirely to government guidance amounts to "complete passivity" as an existential-safety strategy for highly advanced AI systems. ▪ Establish and publicize a whistleblower policy. There is no clear governance structure or whistleblowing channel, even absent reported incidents.
xAI		▪ Build a substantial safety team and engage with existential safety. There is no evidence of a meaningful safety team or any ongoing engagement with existential-safety concerns. ▪ Broaden dangerous-capability evaluations and link thresholds to binding mitigations. Evaluations have gaping holes (no AI R&D data), and no procedure connects threshold breaches to deployment decisions, making thresholds effectively non-binding. ▪ Broaden dangerous-capability evaluations to include important fields such as AI R&D and link deployment decisions to the safety framework, and thresholds to binding mitigations in the safety framework.
DeepSeek		▪ Publish a full safety framework and governance structure. The company has no published framework; its rating largely reflects the Chinese regulatory environment rather than independent safety leadership. ▪ Move beyond passive deference to regulation toward proactive safety research. Deferring entirely to government guidance amounts to "complete passivity" as an existential-safety strategy for highly advanced AI systems. ▪ Establish and publicize a whistleblower policy. There is no clear governance structure or whistleblowing channel, even absent reported incidents.
Mistral		▪ Publish a full safety framework and governance structure. ▪ Engage substantively with existential safety. Leadership consistently downplays — and at times dismisses — frontier risk rather than articulating any control or alignment strategy. ▪ Improve weak safety benchmark performance.

1.3 Methodology

Index Structure: The Summer 2026 Index evaluates nine leading AI companies on 37 indicators spanning six critical domains. The eight companies include Anthropic, OpenAI, Google DeepMind, xAI, Z.ai, Meta, DeepSeek, Alibaba Cloud, Mistral. The indicators are listed below, and more detailed definitions can be found in Section 3.1.

Risk Assessment

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

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

Safety Frameworks

Risk Identification
Risk Analysis and Evaluation
Risk Treatment
Risk Governance

Existential Safety

Existential Safety Strategy
Internal Monitoring and Control Interventions
Technical AI Safety Research
Supporting External Safety Research

Governance & Accountability

Company Structure & Mandate
Whistleblowing Protection
- Whistleblowing Policy Transparency
- Whistleblowing Policy Quality Analysis
- Reporting Culture & Whistleblowing Track Record

Current Harms

Safety Performance
- Stanford's HELM Safety Benchmark
- Stanford's HELM AIR Benchmark
- TrustLLM Benchmark
- Center for AI Safety Benchmarks

Digital Responsibility
- Protecting Safeguards from Fine-tuning
- Watermarking
- User Privacy

Major Safety Incidents & Response
Military Use of AI

Information Sharing

Technical Specifications
- System Prompt Transparency
- Behavior Specification Transparency

Voluntary Commitment
- G7 Hiroshima AI Process Reporting
- EU General-Purpose AI Code of Practice
- Frontier AI Safety Commitments (AI Seoul Summit, 2024)
- FLI AI Safety Index Survey Engagement
- Endorsement of the Oct. 2025 Superintelligence Statement

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

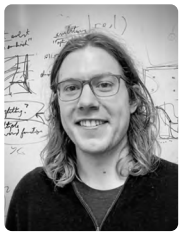
Public Policy
- Policy Engagement on AI Safety Regulations

Data Collection: The Index collected evidence up until June 3, 2026, combining publicly available materials—including model cards, research papers, and benchmark results—with responses from a targeted company survey designed to address specific transparency gaps in the industry, such as transparency on whistleblower protections and external model evaluations. Anthropic, OpenAI, Google DeepMind, Meta and Z.ai have submitted their survey responses. The complete evidence base is documented in Appendix A and Appendix B.

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

1.4 Independent Review Panel

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



David Krueger

David Krueger is an Assistant Professor in Robust, Reasoning, and Responsible AI at the University of Montreal and a Core Academic Member at Mila, the Quebec Artificial Intelligence Institute, and the founder of Evitable. His work focuses on AI alignment, safety, and existential risks from advanced AI. He was a founding Research Director at the UK AI Security Institute and initiated the landmark CAIS Statement on AI Risk.

Sharon Li

Sharon Li is an Associate Professor in the Department of Computer Sciences at the University of Wisconsin-Madison. Her research focuses on algorithmic and theoretical foundations of safe and reliable AI, addressing challenges in both model development and deployment in the open world. She serves as the Program Chair for ICML 2026. Her awards include a Sloan Fellowship (2025), NSF CAREER Award (2023), MIT Innovators Under 35 Award (2023), Forbes 30under30 in Science (2020), and “Innovator of the Year 2023” (MIT Technology Review). She won the Outstanding Paper Award at NeurIPS 2022 and ICLR 2022.

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, a global youthled organization advocating for the ethical regulation of AI. Under her leadership, Encode has mobilized thousands of young people to address challenges like algorithmic bias and AI accountability. She was featured on TIME’s inaugural list of the 100 most influential people in AI.

Stuart Russell

Stuart Russell is a Professor of Computer Science at the University of California at Berkeley and Director of the Center for Human-Compatible AI and the Kavli Center for Ethics, Science, and the Public. He is a member of the National Academy of Engineering and a Fellow of the Royal Society. He is a recipient of the IJCAI Computers and Thought Award, the IJCAI Research Excellence Award, and the ACM Allen Newell Award. In 2021 he received the OBE from Her Majesty Queen Elizabeth and gave the BBC Reith Lectures. He coauthored the standard textbook for AI, which is used in over 1500 universities in 135 countries.

Robert Trager

Robert F. Trager is Director of the Oxford Martin AI Governance Initiative, International Governance Lead at the Centre for the Governance of AI, and Senior Researcher in the Department of Engineering Science at the University of Oxford. He is a recognized expert in the international governance of emerging technologies, diplomatic practice, institutional design, and technology regulation. He regularly advises government and industry leaders on these topics.

Yi Zeng

Yi Zeng is a Wu Yuzhang Chair Professor and Ph.D. Supervisor at the Gaoling School of AI, Renmin University of China, the Founding Dean of Beijing Institute of AI Safety and Governance (Beijing-AISI), the Director of the Beijing Key Laboratory of Safe AI and Superalignment, the Chair of the Mind Computing Technical Committee of the Chinese Association for Artificial Intelligence (CAAI), and Co-Chair of the AI Committee of the World Internet Conference (WIC).He serves on the United Nations Advisory Body on AI and the UNESCO Ad Hoc Expert Group on AI Ethics. He has been named one of TIME's 100 Most Influential People in AI.

2 Introduction

The first half of 2026 marked a turning point in capabilities of advanced AI systems. They have increasingly saturated capability benchmarks and crossed thresholds that companies themselves judged too consequential to release without robust restriction. Anthropic's Mythos system demonstrated an unprecedented ability to autonomously discover and exploit previously unknown software vulnerabilities, identifying thousands of zero-day vulnerabilities across virtually every major operating system and browser, and writing working exploits for them without human guidance, at costs as low as under $50 per run (Carlini et al., 2026). When a safeguarded successor, Fable 5, was finally released to the public in June, the U.S. Commerce Department ordered it suspended within days under an export-control directive citing national security concerns (Anthropic, 2026a). Similar restrictions also applied later to the latest OpenAI Sol model (OpenAI, 2026). Leaders across finance, technology, and government throughout the world were alarmed not by what such systems might one day achieve, but by what they had already been able to do.

The capability increase has been matched by a sharp escalation in real-world harm. Wrongful-death litigation against major AI companies has multiplied. Leading companies including OpenAI and Google have been sued over a growing wave of wrongful-death claims alleging their chatbots encouraged or affirmed users' self-harm in the period leading up to their suicides (Elias, 2026; Bhuiyan, 2025). In addition, AI-enabled targeting has been featured prominently in the U.S. military campaign against Iran, raising acute questions of accountability after strikes that caused mass civilian casualties (De Luce et al., 2026). Anthropic had earlier refused the Pentagon's demand for unrestricted use of its model, forfeiting the contract over concerns about autonomous weapons and domestic surveillance (Hays, 2026); however, when its CEO, Dario Amodei, was pressed on the company's possible role in one such strike, an attack that reportedly killed around 120 children, he conceded he did not know how the system had been used, while maintaining that the use "doesn't even violate our red lines" because a human retained the final decision (Manson & Chang, 2026).

Even some of the companies driving this acceleration have begun, at least rhetorically, to call for restraint. Within days of one another, Anthropic and OpenAI each announced that they endorse the option to slow or pause frontier development, and Google DeepMind's chief executive has said he would support a pause as well (Altman and Pachocki, 2026; Anthropic, 2026b; World Economic Forum, 2026). Yet every such call is heavily conditioned: it would take effect only if multiple well-resourced AI companies across multiple countries, including the United States and China, agreed to stop simultaneously under identical and verifiable terms, with no company committing to move first. Moreover, the proposals also leave their core terms undefined, including what would trigger a pause, what would lift it, and who would adjudicate.

In response to this growing urgency, the AI Safety Index—developed by the Future of Life Institute together with an independent panel of experts in AI safety, governance, and technical evaluation—offers a biannual and independent assessment of how responsibly the world's leading AI companies are developing and deploying advanced systems. The Index evaluates company safety practices across six domains, from risk-management frameworks to pre-deployment evaluations, and from internal governance to external information sharing. In this fourth edition, it does more than rate current behaviors: by holding its indicators largely stable across iterations, the report tracks how each company's practices evolve over time, making both genuine progress and persistent gaps visible. By presenting results in a format accessible to specialists and general audiences alike, the Index turns scattered commitments and claims into a transparent, evidence-based, and comparable picture, at a moment when some argue that a widening gap between capability, real-world harm, and the credibility of corporate safety pledges demand external verification rather than self-attestation more than ever.

3 Methodology

The AI Safety Index evaluates and grades the safety practices from AI companies in four steps: indicator selection, company selection, evidence collection, and grading.

3.1 Indicator Selection

To closely examine AI companies' safety practices throughout the lifecycle, we have selected 37 indicators: 35 from the Winter 2025 version, 2 newly added.

The domains capture different aspects of responsible AI development and deployment, including Risk Assessment, Current Harms, Safety Framework, Existential Risk Strategy, Governance and Accountability, as well as Information Sharing and Public Messaging, echoing principles embedded in regulatory obligations and voluntary-commitment frameworks including the EU AI Code of Practice and the G7 Hiroshima Process. In particular, the Index highlights the existential risk strategy—a dimension not explicitly addressed in leading governance frameworks—because proactive planning for existential risk has become a pressing need, as emphasized by leading AI technical researchers and governance experts, including Bengio et al. (2024).

Two new indicators were added to the Current Harms domain to more comprehensively track companies' real-world harm record and their involvement in high-stakes settings. The first, Major Safety Incidents & Response, assesses how companies handle harms caused by or allegedly caused by their AI systems; the second, Military Use of AI, examines how they govern the integration of their systems into high-stakes contexts involving life-and-death decisions.

Risk Assessment

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

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

Current Harms

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

Safety Performance	Stanford's HELM Safety Benchmark	Evaluates how language models perform on key safety metrics like robustness, fairness, and resistance to harmful behavior.
	Stanford's HELM AIR Benchmark	Measures AI model safety and security on benchmark aligned with emerging government regulations and company policies.
	TrustLLM Benchmark	Assesses a model's trustworthiness across dimensions such as safety, ethics, and alignment with human values and expectations.
	Center for AI Safety Benchmarks	Measures AI safety behaviors including resistance to misuse, appropriate refusals, calibration accuracy, honesty under pressure, and ethical restraint in scenarios.
Digital Responsibility	Protecting Safeguards from Fine-tuning	Evaluates whether AI providers implement protections that prevent fine-tuning from disabling important safety mechanisms or filters.
	Watermarking	Assess whether AI outputs are marked in a detectable way to help track origin and reduce misinformation or misuse.
	User Privacy	Measures the degree to which an AI company protects user data from extraction, exposure, or inappropriate use by models.
Major Safety Incidents & Response		The indicator documents the record of publicly reported safety incidents involving the company's AI products where serious potential or actual harm occurred, and evaluates how the company responded in each case.
Military Use of AI		The extent to which a company supplies, partners with, or develops AI systems for military and defense applications (e.g., targeting, surveillance, autonomous weapons, or battlefield decision support), and the policy, transparency and safeguards governing those engagements.

Safety Frameworks

This domain evaluates the companies' published safety frameworks for frontier AI development and deployment from a risk management perspective. The analysis follows the taxonomy and indicator structure developed by the non-profit research organization SaferAI.

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

Existential Safety

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

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

Governance & Accountability

This domain evaluates how openly companies share technical, safety, and governance information, and how their public and legislative messaging align with responsible AI governance

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

Information Sharing

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

Technical Specifications	System Prompt Transparency	Assesses whether companies publicly disclose the actual system prompts used in their deployed AI models, including version histories and design rationales.
	Behavior Specification Transparency	Evaluates if developers publish detailed and up-to-date documentation explaining their models' intended behavior, values, and decision-making logic across diverse scenarios.
Voluntary Cooperation	G7 Hiroshima AI Process Reporting	Tracks whether companies submitted detailed safety and governance disclosures to the G7 Hiroshima AI Process, reflecting their commitment to transparency.
	EU General-Purpose AI Code of Practice	Demonstrates AI companies' voluntary compliance with EU AI Act General-Purpose AI (GPAI) obligations by signing the non-binding guidelines.
	Frontier AI Safety Commitments (AI Seoul Summit, 2024)	Measures adherence to voluntary pledges by leading AI companies to develop safety frameworks for evaluating and managing severe AI risks.
	FLI AI Safety Index Survey Engagement	Reports which companies voluntarily completed and submitted FLI's detailed safety survey to supplement publicly available information.
	Endorsement of the Oct. 2025 Superintelligence Statement	Indicates whether a company has endorsed calls to prohibit superintelligence development until broad scientific consensus confirms safety and controllability.
Risks & Incidents	Serious Incident Reporting & Government Notifications	Evaluates public commitments, frameworks, and track records around reporting serious AI-related incidents to governments and peers.
	Extreme-Risk Transparency & Engagement	Measures whether company leaders publicly acknowledge catastrophic AI risks and proactively communicate those concerns to external audiences.
Public Policy	Policy Engagement on AI Safety Regulations	Tracks company involvement in shaping AI safety laws through public statements, consultations, testimony, and participation in regulatory coalitions.

3.2 Company Selection

The Index is primarily focused on companies that have deployed the most highly capable models currently available, or those that have previously done so and continue to actively invest in the development and deployment of new frontier systems. In addition, we include at least one company from each of the three leading AI continents (North America, Europe and Asia), to retain in any company that met our other inclusion criteria for the previous index. We plan to broaden coverage to additional regions and companies as the frontier landscape evolves.

Combining insights from leaderboard rankings from Arena obtained on April 11, 2026, investigation on continued investment into frontier AI systems development and deployment, and regional representation, the companies selected for this iteration are: Alibaba Cloud, Anthropic, DeepSeek, Google DeepMind, Meta, Mistral, OpenAI, xAI, Z.ai.[1] Although DeepSeek at the time of model selection did not offer a model at the top-10 capability frontier, we are keeping it in the Index for one additional iteration in recognition of its sustained investment toward superintelligence-level research.[2]

We are grateful to the Arena team for providing historical leaderboard data, which enabled us to assess not only companies' current model capabilities but also their sustained presence at the frontier over time.

3.3 Related Work

Related Work: Several notable related efforts that drive transparency and accountability within the industry continue to inspire and complement the AI Safety Index. The most comprehensive of these efforts include SaferAI's in-depth analysis and ranking of AI companies' public safety frameworks (most recently updated as of July 2026), and two projects by Zach Stein-Perlman—AILabWatch.org (most recently updated as of September 15, 2025) and AISafetyClaims.org (most recently updated as of September 1, 2025)—which have offered inspirations to provide detailed and technical evaluations of how leading AI companies work to avert catastrophic risks from advanced AI. Complementing these, the OECD report published in September 2025 synthesizes disclosures submitted through the G7's voluntary reporting framework and offers one of the first comparative, policy-grounded views of companies' governance and risk-management practices (Perset and Fialho Esposito, 2025). Earlier efforts include the Foundation Model Transparency Index in October 2023 and May 2024 published by Stanford Center for Research and Foundational Models (CRFM), which provides an empirical baseline for model transparency across the ecosystem.

Incorporated Work: Where appropriate, the Summer 2026 Index incorporates existing comparative analysis led by credible research institutions.

In the Safety Frameworks domain, the Index draws on the indicator set developed by SaferAI for its in-depth assessment of companies' published safety frameworks, as well as referencing its underlying evidence base for building the Safety Index's evidence, which the SaferAI team kindly provided — while leaving all scoring to the independent reviewers convened by FLI. SaferAI is a leading governance and research non-profit with significant expertise in AI risk management.

The Index further integrates AILabWatch.org's tracking of technical AI safety research within the Existential Safety domain and complements it in two ways: by adding research published after the tracker's most recent update, and by incorporating safety-relevant research from companies not included in AILabWatch's coverage.

Our research on the quality of companies' whistleblowing policies in the 'Governance & Accountability' domain

1 The archived leaderboard on April 11, 2026 can be retrieved at this link: https://archive.is/OOYTr. The companies are ordered in alphabetical order.

2 DeepSeek released its V4 model after the company selection cut-off date of the AI Safety Index Summer 2026.

was enabled through support from The AI Whistleblower Initiative, a non-profit supporting individuals working at the frontier of AI who want to flag risks.

The 'Current Harms' domain evaluates flagship model performance on leading safety benchmarks, including the TrustLLM benchmark, the HELM AIR-Bench and HELM Safety benchmarks by Stanford's CRFM, and the Safety Index benchmarks curated by the Center of AI Safety (CAIS) AI Dashboard.

3.4 Evidence Collection

The evidence collected for this iteration of the Index covers information up until June 3, 2026, drawing from publicly available information and a dedicated company survey for additional voluntary disclosures. Throughout the data collection process, the FLI team aimed to minimize bias and ensure a fair evaluation by applying consistent search protocols and evidence standards across companies.

To ensure fair evaluation across companies operating in different jurisdictions, and to familiarize panelists who may be less acquainted with recent legislative developments in China, this iteration retains a concise, structured section explaining how China's regulatory system—across binding national laws, local regulations, voluntary technical standards, draft instruments, and policy guidance—shapes company behavior and disclosure practices. This addition enables reviewers to interpret Chinese companies' evidence within the regulatory environment they operate in, rather than through assumptions derived from US and Europe contexts that emphasize voluntary self-governance and public documentation. By integrating this regulatory mapping into each relevant domain, the Index aims to improve cross-jurisdictional comparability and reduce systematic bias in grading.

In addition, this iteration expands the structured mapping beyond the EU AI Code of Practice to include U.S. state legislation that is either already in force or set to take effect, including California's SB 53, the New York RAISE Act, and Illinois SB 315. For each domain, we identify the most relevant provisions and present them as a baseline reference for the voluntary and legal obligations many of the included companies currently face. This mapping is provided solely as contextual material to help reviewers situate the indicators within emerging governance expectations; it does not prescribe grading thresholds or function as an official rubric. Instead, panelists are encouraged to apply their own expert judgment, drawing on the mapping as one of several reference points when interpreting companies' safety practices — particularly as companies navigate both compliance expectations and their own frontier-model development ambitions.

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

Company survey: To supplement public information, FLI created a 35-question survey that addresses current gaps in voluntary disclosures. The survey was sent out via e-mail on April 20, 2026. The survey can be reviewed in full in Appendix B.

While the survey remains largely consistent with prior editions to allow comparison over time, this iteration refines its treatment of external evaluations. Question 17 and Question 18 asks respondents to clarify the nature of each external evaluator's relationship with the company, including financial ties, governance ties, and material commercial dependency. This change allows the Index to capture the full range of external evaluation activity while separately assessing the degree of evaluator independence.

We received survey responses from five of the nine companies (Anthropic, Google DeepMind, Meta, OpenAI, and Z.ai). xAI failed to submit for the first time since the first edition of AI Safety Index, while Alibaba Cloud, DeepSeek, and Mistral did not submit a response. Full survey responses are attached in Appendix B.

Grading Sheets: The evidence collected for this edition of the Index was organized into the grading sheets presented in Appendix A. These sheets are divided across six domains and provide company-specific information for each of the indicators included in the current edition. For every indicator, the grading sheets outline its scope, explain the rationale for its inclusion, and reference relevant literature with hyperlinks where appropriate. We prioritized primary sources directly from companies over secondary reporting wherever possible. Investigative journalism played an important role by surfacing practices that companies have not publicly disclosed. Survey responses submitted by companies were incorporated and clearly highlighted within the relevant indicators. Each domain also includes a concise description of the corresponding Chinese regulatory environment. Where applicable, indicators are mapped to commitments in the EU AI Code of Practice and provisions in relevant U.S. state legislations to help situate them within emerging governance expectations.

3.5 Grading

The grading process was designed to ensure an impartial and qualified evaluation of the companies' performance across the selected indicators, based on expertise of individual reviewers in relevant fields. It features a review panel of distinguished independent experts who assess the company-specific evidence for their assigned indicators and assign domain-level grades that represent companies' performance within these domains.

Review Panel: To ensure that the Index scores rest upon authoritative judgements, FLI selected a group of seven leading independent experts to grade company performance on the set of indicators. Panel members were selected for their domain expertise and absence of conflicts of interest. Because the Index spans technical AI safety, governance, and policy, the panel brings together specialists across these areas and reflects broad geographic diversity. The panel thus features both renowned machine learning professors who specialize in alignment and control, and governance experts from the academic and non-profit sectors. The composition of the panel remained largely unchanged from the previous edition. We are grateful to Robert Trager for joining the panel as its new member. The review panel is introduced at the beginning of this document.

Grading Phase: Grading sheets and survey results were shared with the review panel for evaluation on June 3, 2026. After reviewing the evidence, reviewers assigned letter grades to each company per domain. For each grade assigned to individual companies, reviewers could provide brief justifications and recommendations. They were also able to provide domain-level comments when feedback applied to multiple firms or to explain their judgments. Not every reviewer graded every domain, but experts were assigned domains relevant to their area of expertise. Importantly, no fixed weighting was imposed across indicators within a domain. This approach allowed expert reviewers to apply their judgment in emphasizing aspects they deemed most critical. The grading sheets provided to reviewers further contained grading scales based on absolute performance standards rather than relative rankings, ensuring consistent expectations regardless of company size or geography. Final domain scores were calculated by averaging all reviewer grades for that domain. Overall grades were then derived by averaging the domain-level scores.

3.6 Limitations

Information Availability and Verification

Our evaluation relies primarily on public information, which creates fundamental constraints. Companies control what they disclose, despite occasional cases of whistleblowing, making it difficult to distinguish between poor transparency and poor strategy and implementation. We designed indicators around these transparency constraints, focusing where meaningful differences between companies were identifiable. For example, we cannot assess critical practices such as cybersecurity investments to protect model weights, as this information is rarely disclosed publicly but we instead look at how companies assess cybersecurity-related risks with their frontier AI systems.

The indicators represent a subset of important practices for which meaningful evidence exists, but it does not comprehensively cover all safety dimensions. Furthermore, we cannot independently verify individual company claims and must assume official reports are truthful, which constitutes a significant limitation given the high stakes involved.

Current Harm Scope

The Current Harms domain currently focuses on observable, measurable outcomes — primarily model performance on established safety benchmarks and the robustness of implemented safeguards against adversarial attacks. As one reviewer noted, this is "not a fully representative assessment of current harm." Real-world harms increasingly extend well beyond what standardized benchmarks capture — including AI-associated psychosis and self-harm, wrongful deaths, erosion of users' cognitive independence, downstream effects on education systems, concentration of power, and environmental costs — many of which are difficult to quantify, attribute, or compare consistently across companies. Our reliance on benchmarks and documented incidents therefore likely understates the true scope of harm.

We should treat benchmark performance as a floor rather than a comprehensive picture, and we intend to broaden this domain's coverage of harder-to-measure harms in future iterations.

Methodological Constraints

Our focus on observable, documentable practices may undervalue crucial but hard-to-measure factors such as safety culture. Additionally, while we seek to diversify the grading panel with specialized expertise and geolocation focus, it cannot encompass all relevant domains across the companies that we review. Panelists' backgrounds inevitably shape their judgments, and there is an inherent tension between allowing experts to exercise domain-specific discretion in weighting indicators and maintaining full consistency across panelists and domains.

Moving Forward

We seek to address these limitations through continued refinement of our methods and closer engagement with policymakers, researchers, and practitioners who rely on the Index. Feedback from regulators and policy professionals is particularly valuable in helping us identify where clearer disclosure expectations, stronger reporting norms, or more precise indicator design would make the Index more actionable for real-world governance needs.

We will continue to document our sources, assumptions, and reviewer materials transparently, and we welcome constructive guidance on how to better incorporate hard-to-evaluate practices, reduce ambiguity in evidence interpretation, and strengthen cross-jurisdictional comparability. We encourage readers to share suggestions at sabina@futureoflife.org and policy@futureoflife.org. We remain committed to advancing the Index with each iteration.

4 Results

Overall Rankings: Anthropic leads with a **C+ (2.66)**, followed by OpenAI **(C+, 2.28)** and Google DeepMind **(C, 2.01)**. The next tier of companies consist of Meta **(D+, 1.32)**, Z.ai **(D-, 0.88)**, and Alibaba Cloud **(D-, 0.87)**. Three companies, xAI **(0.65)**, DeepSeek **(0.47)**, and Mistral **(0.33)** received failing grades. Notably, like the last iteration, no company scored above a C+, underscoring that even the strongest performers remain far from meeting adequate safety expectations.

	Anthropic	OpenAI	Google DeepMind	Meta	Z.ai	Alibaba Cloud	xAI	DeepSeek	Mistral
Overall Grade	C+	C	C	D+	D-	D-	F	F	F
Score	2.66	2.28	2.01	1.32	0.88	0.87	0.65	0.47	0.33
Grade Trend (Winter 25)	C+	C+ ▼	C	D ▲	D ▼	D-	D ▼	D ▼	N/A
Risk Assessment (6 indicators)	C+	C+	C+	D+	F	F	D-	F	F
Current Harms (9 indicators)	B-	C	C	D-	C-	C-	F	D-	F
Safety Frameworks (4 indicators)	B-	C+	C	C-	D-	D-	D	F	F
Existential Safety (4 indicators)	D+	D+	D	F	F	F	F	F	F
Governance & Accountability (4 indicators)	B	C	C-	D+	F	D-	F	F	F
Information Sharing (10 indicators)	B+	B-	B-	D+	D	D	D	D-	D-
Survey Responses	✓	✓	✓	✓	✓	✗	✗	✗	✗

Grading: Uses the US GPA system for grade boundaries: A+, A, A-, B+, [...], F letter values corresponding to numerical values 4.3, 4.0, 3.7, 3.3, [...], 0.

4.1 Key Findings

- **Anthropic, OpenAI, and Google DeepMind stay on top.** Anthropic again earns the highest overall grade and leads five of six domains via relatively strong transparency, a comparatively established safety framework, technical research, and governance. OpenAI now leads in Risk Assessment on the strength of a broader evaluation suite and diverse engagement with external testing.
- **Meta improves and xAI deteriorates:** Meta improved from 6th to 4th place, while xAI dropped from 4th to 7th place.
- **European dissonance:** Although the European Union is a leader in AI safety regulation, the top European AI company Mistral scored dead last on safety.
- **Inadequate safety is a global problem, not a regional one.** Three companies receive failing grades, one each from the US (xAI), China (DeepSeek), and Europe (Mistral).
- **Reviewers flagged the industry's pivot to military AI use as an emerging current harm risk.** From 2024 to

2026, companies including Anthropic, OpenAI, Google DeepMind, and Meta that previously banned military applications gradually reversed course, joining xAI and Mistral in actively seeking defense partnerships. Despite their limits on domestic surveillance and autonomous weapons, Anthropic drew criticism from the review panel for "questionable military engagements," including a reported link to the Minab school strike that caused mass civilian deaths. Leading Chinese firms, meanwhile, face U.S. allegations of military ties that Alibaba Cloud and Z.ai deny.

- **Even industry leaders in safety practices are retreating from prior commitments.** Anthropic, OpenAI, Google DeepMind, and Meta have weakened or voided pledges to pause unilaterally if redlines are approached, some citing competitor-contingent conditions. Reviewers call this "moving goalpost" and argue that it has "undermined safety frameworks across the board".
- **Existential Safety is the weakest domain industry-wide.** No company exceeds C-; most score D or below. Constructive attempts exist, such as Anthropic's constitutional classifiers, OpenAI's call for governance institutions, Google DeepMind's monitoring commitments, and Meta's loss-of-control provisions, but are judged by panelists to be "entirely inadequate." Dominant paradigms such as interpretability and Chain-of-Thought (CoT) monitorability are questioned because "detection is not prevention."
- **Safety rhetoric outpaces revealed behavior.** Across Google DeepMind, OpenAI, and xAI, leadership's reassuring public messaging diverges from commercial conduct and legislative stance, making stated commitments an unreliable proxy for actual safety practice.
- **Companies are publishing and updating safety frameworks, but these frameworks have weak teeth.** As US/EU compliance deadlines near, Anthropic, OpenAI, Google DeepMind, Meta, and xAI published and updated fuller frameworks — yet they sometimes lack quantitative thresholds, genuinely independent audits, and clear decision authority.

4.2 Company Progress Highlights and Key Recommendations

Company	*Progress Highlights*	*Key Recommendations*
Anthropic	▪ Comparatively detailed safety framework with commitments for third-party audits. ▪ Solid evaluations of autonomous R&D and scheming/misalignment capabilities of frontier models with strong elicitation. ▪ Continuous industry-leading transparency with both published model specs and system prompts.	▪ Reverse the RSP 3.0 walk-back on pause commitments and restore credibility of commitments. ▪ Replace qualitative thresholds with quantitative and risk-tied ones. ▪ Treat prevention as seriously as interpretability/detection. ▪ Establish stronger safeguards for military use of its AI systems.
OpenAI	▪ Strong external testing and comparatively broad risk assessment. ▪ Called for global governance institutions to slow development when needed. ▪ Regular reports documenting their disruption of malicious uses of their AI systems.	▪ Remove leadership's ability to override the Safety Advisory Group. ▪ Make safety-framework thresholds measurable, risk-tiered, and externally enforceable, with commitment to notify authorities when risk thresholds are crossed. ▪ Evaluate internal-deployment risks before broad internal use rather than after. ▪ Align public-policy positions with stated safety commitments.
Google DeepMind	▪ Updated Frontier Safety Framework adding manipulation, misalignment, and internal-deployment coverage. ▪ Strong watermark protection.	▪ Establish clear decision-making authority, an executive risk officer, and independent audit. It remains unclear which internal body can halt deployment independently of executive leadership. ▪ Make safety-framework thresholds measurable and risk-tiered. ▪ Reverse the backsliding on pause commitments. ▪ Align public-policy positions with stated safety commitments from leadership.

Company	Progress Highlights	Key Recommendations
Meta	▪ Published safety framework with more details of risk identification and threat modeling. ▪ Bug bounties cover catastrophic risk factors.	▪ Strengthen whistleblower protections and align culture with policy. The whistleblowing policy quality scores are reasonable but undermined by active enforcement of a non-disparagement agreement and other suppression of dissent. ▪ Make safety-framework thresholds measurable and risk-tiered. ▪ Establish auditing mechanisms for the safety framework.
Z.ai	▪ More transparent than its Chinese peers, with some proactive safety research built into products. ▪ Some meaningful incident-response infrastructure and internal-deployment threat mitigation.	▪ Publish a full safety framework and governance structure. The company has no published framework; its rating largely reflects the Chinese regulatory environment rather than independent safety leadership. ▪ Move beyond passive deference to regulation toward proactive safety research. Deferring entirely to government guidance amounts to "complete passivity" as an existential-safety strategy for highly advanced AI systems. ▪ Establish and publicize a whistleblower policy. There is no clear governance structure or whistleblowing channel, even absent reported incidents.
Alibaba Cloud	▪ Stronger-than-expected benchmark performance with more transparent disclosure of misalignment propensity than its peers. ▪ Two-layered safety strategy spanning the Qwen model-level team and the Alibaba security team.	▪ Publish a full safety framework and governance structure. The company has no published framework; its rating largely reflects the Chinese regulatory environment rather than independent safety leadership. ▪ Move beyond passive deference to regulation toward proactive safety research. Deferring entirely to government guidance amounts to "complete passivity" as an existential-safety strategy for highly advanced AI systems. ▪ Establish and publicize a whistleblower policy. There is no clear governance structure or whistleblowing channel, even absent reported incidents.
xAI		▪ Build a substantial safety team and engage with existential safety. There is no evidence of a meaningful safety team or any ongoing engagement with existential-safety concerns. ▪ Broaden dangerous-capability evaluations and link thresholds to binding mitigations. Evaluations have gaping holes (no AI R&D data), and no procedure connects threshold breaches to deployment decisions, making thresholds effectively non-binding. ▪ Broaden dangerous-capability evaluations to include important fields such as AI R&D and link deployment decisions to the safety framework, and thresholds to binding mitigations in the safety framework.
DeepSeek		▪ Publish a full safety framework and governance structure. The company has no published framework; its rating largely reflects the Chinese regulatory environment rather than independent safety leadership. ▪ Move beyond passive deference to regulation toward proactive safety research. Deferring entirely to government guidance amounts to "complete passivity" as an existential-safety strategy for highly advanced AI systems. ▪ Establish and publicize a whistleblower policy. There is no clear governance structure or whistleblowing channel, even absent reported incidents.
Mistral		▪ Publish a full safety framework and governance structure. ▪ Engage substantively with existential safety. Leadership consistently downplays — and at times dismisses — frontier risk rather than articulating any control or alignment strategy. ▪ Improve weak safety benchmark performance.

4.3 Domain-level findings

Risk Assessment

	Anthropic	OpenAI	Google DeepMind	Meta	Z.ai	Alibaba Cloud	xAI	DeepSeek	Mistral
Domain Grade	C+	C+	C+	D+	F	F	D-	F	F

Companies are divided into clear tiers in this domain, with Anthropic, OpenAI, and Google DeepMind continuing to lead, Meta a step behind, and the remaining companies offering little of substance to assess. The leaders have maintained documented evaluation processes in relation to their safety frameworks, recording structured elicitation through testing on helpful-only model variants and scaffolding, established external testing partnerships, and bug-bounty programs. Two reviewers singled out OpenAI to be the broadest in terms of risk coverage, while one reviewer credited Anthropic with strong evaluations for autonomous R&D, scheming, and misalignment. Google DeepMind sits slightly behind on the granularity of its model-card reporting but remains far ahead of the remaining companies. Meta has conducted more internal and external evaluations, despite the omission for manipulation evaluations. xAI's assessment contains what one reviewer called "gaping holes," including no AI R&D data and weak elicitation, against a backdrop of a regressing Grok 4 model with concerning propensities. The remaining AI companies offer little to no substance to assess.

Yet even for the strongest players, reviewers identified structural gaps that begin with how risk itself is defined. One panelist found an "overly narrow definition of dangerous capabilities, accompanied with no discussion of how risks are chosen," and warned that the prevailing focus is "out of touch with the reality of risks (including death)." This narrowness invites a "safety-washing" critique — companies "pushing their product as safer than alternatives" when, for the largest risks, the differences are marginal.

In addition, these weaknesses converge on three near-universal omissions. First, no company conducts human uplift trials for the flagship models selected for this index, flagged by one reviewer as "a big threat vector." Second, almost no company secures genuinely independent review of its safety evaluations — neither independent audit of internal evaluations nor truly independent external assessment. Reviewers noted that Anthropic has "no independent audit and minimal external footprint" and that DeepMind has "no formal audit." Where external testing does exist, its independence is constrained by developer-offered financial compensation, including DeepMind and OpenAI, and by limited disclosure of governance independence and publication freedom by other companies. Third, internal-deployment risks remain under-examined: OpenAI is the only company to disclose the relationship between its external evaluations and internal deployment, and it indicates that external testing "was only completed after some period of broad internal use" — leaving the internal-deployment window itself largely unassessed.

Safety Frameworks

	Anthropic	OpenAI	Google DeepMind	Meta	Z.ai	Alibaba Cloud	xAI	DeepSeek	Mistral
Domain Grade	B-	C+	C	C-	D-	D-	D	F	F

Five out of nine companies, including Anthropic, OpenAI, Google DeepMind, Meta, and xAI, have published and updated safety frameworks, whereas DeepSeek, Z.ai, Alibaba Cloud, and Mistral have none and receive only an increase in score for "signing" relevant safety commitments, which one reviewer dismissed as "generic and used only as a reference." Among those that have published, quality is concentrated at the top. Anthropic's framework was judged the most detailed with "definitions, thresholds, risk tiers, and threat modeling," as well as accountability framework such as Responsible Scaling Officer (RSO), third-party procedural compliance and external reviews of risk reports. OpenAI and DeepMind follow, the latter credited for sound RAND-based security-level mapping. Meta was recognized for "strong threat modeling" but flagged as inadequate in Chem and Bio risks, "despite this being a severe threat vector," while xAI offers "some useful threat modeling and risk-identification thresholds" while lacking provisions on governance. Z.ai and Alibaba Cloud received a small raise in the grade due to dedicated safety teams, as recognized by reviewers.

However, underlying the leaders' commitments is a concern about durability. One reviewer judged that "Anthropic's retreat on its safety framework has undermined safety frameworks across the board" by rolling back its commitment to a unilateral pause towards competitor-contingent "consideration" for pause. The pattern extends beyond Anthropic: one reviewer noted that OpenAI has rolled back commitments through its updated Preparedness Framework, which "leaves room for adjustments on requirements contingent upon competitor behaviors," that Google DeepMind has effectively voided the pause commitments specified in earlier versions of its safety framework, and that Meta has also backslided on pause commitments. This points to a corrosive structural feature reviewers identified as applying to "all the developers," that "strict application is not required when competitors do not apply similar practices," a conditional posture that incentivizes a collective race to the bottom.

Beyond commitment durability, existing safety frameworks share certain structural gaps across the industry. For one, thresholds are rarely quantitative, a weakness persisting since earlier iterations, leaving little clarity, and little opportunity for public scrutiny, over what specific evidence would trigger stronger safeguards. Multiple reviewers further flagged that binding authority and independent oversight remain weak across published frameworks. OpenAI permits "leadership to override" its Safety Advisory Group; DeepMind has been criticized for "a lack of clear decision-making authority" on frontier AI safety concerns, with no internal or external audit mechanism; Meta's framework "narrows what counts as in-scope risk," leaves "thresholds undisclosed," and specifies "[no] Board power over deployment decisions"; and xAI "lacks a defined decision-making authority, and any advisory, audit, or oversight body," rendering its thresholds "effectively non-binding."

Current Harms

	Anthropic	OpenAI	Google DeepMind	Meta	Z.ai	Alibaba Cloud	xAI	DeepSeek	Mistral
Domain Grade	B-	C	C	D-	C-	C-	F	D-	F

Benchmark performance is the clearest dividing line in this domain, though reviewers cautioned it captures only part of the picture. Anthropic and OpenAI score the strongest, with Anthropic noted for "good benchmark robustness and good hard red lines around issues such as domestic surveillance"; Alibaba Cloud was judged "stronger than expected in terms of benchmark performance" with "transparent disclosure of misalignment propensity"; and Mistral and DeepSeek perform significantly poorly compared to other companies. But reviewers stressed that benchmark scores understate real risk: passing them amounts to clearing "basic functionality," and several reviewers have highlighted that strong scores coexist with serious deployed harms.

The most consequential observations concern documented real-world harms, which pull down the grades of companies with otherwise strong benchmark performance. xAI, which received a failing grade in this domain, was criticized for reported "mass generation of CSAM and Grok being used to nudify people including minors." Reviewers also pointed to the observed histories of OpenAI's and Google DeepMind's products in alleged cases of self-harm and wrongful death: OpenAI in connection with the Adam Raine case, in which the family's lawsuit alleges that "safeguards [were] removed over time," and DeepMind through lawsuits in relation to Character.ai, and its chatbot allegedly "leading to wrongful deaths and self-harm." One reviewer noted Meta as "notorious for kids' safety issues, leading to a $375 million jury finding," as well as sexual-companion-chatbot incidents. Alibaba Cloud's incident involved "agentic behaviors without permission," though reviewers credited Alibaba for having "properly addressed it."

In addition, reviewers have noted the industry's shift towards embracing military applications of advanced AI systems. Reviewers weighed this explicitly only for Anthropic, an early national-security mover whose Pentagon work and reported link to the Minab school strike drew criticism as "questionable military engagements," even as Anthropic held red lines on domestic surveillance and autonomous weapons. More broadly, from 2024 to 2026, companies such as Anthropic, OpenAI, Google DeepMind, and Meta all reversed their earlier bans on military use over the course, joining xAI and Mistral in actively pursuing defense partnerships. xAI is committed to deploying AI systems in the military space "without ideological constraints." Meanwhile, Mistral builds AI applications in defense around European sovereignty while disclaiming corporate responsibility for deployment. DeepSeek, Z.ai, and Alibaba Cloud face U.S. allegations of PLA support, which Z.ai and Alibaba deny publicly.

Existential Safety

	Anthropic	OpenAI	Google DeepMind	Meta	Z.ai	Alibaba Cloud	xAI	DeepSeek	Mistral
Domain Grade	D+	D+	D	F	F	F	F	F	F

Frontier AI companies have been consistently weak in this domain: the leading AI companies are advancing towards transformative AI at unprecedented speeds, yet they lack credible plans to control it. Even the three strongest performers — Anthropic, OpenAI, and Google DeepMind — were found to have "no explicit safety strategy," scoring above the rest only because they have had more substantial track records of research in the

related fields and support for external researchers. Multiple reviewers described Anthropic and OpenAI as "racing towards recursive self-improvement, risking an irreversible loss of control," with one judging Anthropic's mitigation plan "sensible but entirely inadequate." Reviewers have noted positively that Meta has started to engage with frontier risks concerns. For companies including Z.ai, Alibaba Cloud, xAI, DeepSeek, and Mistral, reviewers have noted that these companies have "little evidence of engagement with existential safety concerns" or "little credible strategy."

Reviewers also questioned the dominant technical paradigms themselves. One argued at length that "interpretability as a primary technical safety strategy" is inadequate because "detection is not prevention," because it addresses "most" rather than all problems, and because it "shifts all technical focus toward internals of the model, [therefore] by construction ignoring operational context." One reviewer judged that "alignment as a primary safety strategy" is also inadequate, noting that the largest realized risks, including "psychosis/suicide and cybersecurity," are cases where alignment is arguably "working 'too well'" through sycophancy and malicious compliance. Chain-of-thought monitoring drew parallel skepticism as "an untrustworthy post-hoc fine-tuning method" sold as if it revealed a model's genuine reasoning.

Governance & Accountability

	Anthropic	OpenAI	Google DeepMind	Meta	Z.ai	Alibaba Cloud	xAI	DeepSeek	Mistral
Domain Grade	B	C	C-	D+	F	D-	F	F	F

Only two of nine companies — Anthropic and OpenAI — have published a whistleblower policy, while two others — Meta and Google DeepMind — have shared policy details via the company survey or published whistleblowing-adjacent infrastructure such as a harassment policy with an anonymous hotline. Where these exist, reviewers credited genuine progress: multiple accessible reporting channels including anonymous options, coverage of AI-specific topics such as safety-framework violations and misleading communications to regulators, emerging privacy protections, and some anti-retaliation provisions. They lag, however, in the back half of the process, from investigation standards, to system governance and quality assurance, where even the leaders score at or near zero.

However, whistleblowing policy quality does not translate into perfect track records. For companies including OpenAI, DeepMind, Meta, and Alibaba, reviewers found that "the whistleblowing track record is not consistent with [the] policy" — most sharply at Meta, "brought down by actively enforcing a non-disparagement agreement and other cases of suppressing dissent," alongside DeepMind's "retaliation complaints and employee dissent over military contracts," OpenAI's alleged firing of Ryan Beiermeister, and Alibaba Cloud's documented "retaliation against [a] sexual assault whistleblower."

For the lower tier companies, the assessment turned largely on absence of evidence, and reviewers diverged: some found "not much to evaluate" for DeepSeek, Z.ai, and Mistral, while others recognized that "there is also no reported whistleblower or retaliation incidents," which neither demonstrates robust accountability nor, in most cases, generates documented failures.

Information Sharing & Public Messaging

	Anthropic	OpenAI	Google DeepMind	Meta	Z.ai	Alibaba Cloud	xAI	DeepSeek	Mistral
Domain Grade	B+	B-	B-	D+	D	D	D	D-	D-

Anthropic, OpenAI, and Google DeepMind lead a clear division from the rest of the companies. On technical transparency, Anthropic is the only company sharing both its system prompt and behavior specification, while OpenAI shares a behavior specification only and xAI has released some system prompts but no behavior spec. Every other company discloses neither. Anthropic is also credited with generally positive policy engagement, whereas OpenAI and Google DeepMind adopt a more hostile and at times complicated stance toward AI safety regulations. Reviewers singled out OpenAI's shifting posture in particular — opposing SB 53 and then claiming "reverse federalism," backtracking on Illinois legislation within a month, and maintaining a "connection to super PAC Leading the Future."

The most consistent theme is a gap between leadership rhetoric and corporate conduct. Reviewers identified it most pointedly at Google DeepMind — a divergence between what its CEO Demis Hassabis conveys through public messaging and how Google acts as a political entity, with the company "largely uncooperative on policy" — and at xAI, which is "largely silent on most policy questions aside from Musk's individual support for [SB] 1047," reflecting a "gap between Musk's communications and corporate communications/actions." Despite a "weak disclosure record" and an adversarial stance toward safety regulation, reviewers noticed that Meta has displayed attitude shifts towards risk implications of advanced AI systems after the leadership reshuffle. Of all the companies, Mistral has been the most dismissive of frontier AI risk, consistently downplaying its importance.

5 Conclusion

The first half of 2026 made the central tension of this Index impossible to ignore: capabilities are advancing faster than the governance and technical approaches meant to govern them, and the consequences are no longer hypothetical. Systems have saturated some of the benchmarks built to measure their most dangerous capabilities, while real-world harm has escalated sharply — from a mounting wave of wrongful-death and self-harm litigation against leading AI companies to the integration of advanced AI systems into military decision-making, in at least one case reportedly contributing to mass civilian casualties.

Against these rising stakes, the review panel found that companies have not kept pace; and in several respects, they argued, companies are moving backward. According to the panelists, no industry leader has a credible plan to control the increasingly autonomous and capable systems it is building. The thresholds that are meant to govern critical development and deployment decisions are kept qualitative rather than holding specific and firm. And public safety rhetoric increasingly diverges from commercial and political conduct. But what reviewers considered the most troubling of all is that the leaders racing hardest toward the capability frontier — Anthropic, OpenAI, Google DeepMind, and Meta — have weakened or voided their prior commitments to pause, including replacing unilateral pledges with competitor-contingent conditions. In the panel's assessment, this dynamic has undermined safety frameworks across the board and incentivizes a collective race to the bottom, rather than to the top.

There are, however, what panelists viewed as genuine grounds for action. A growing body of legislation now offers real levers for accountability — including independent audit requirements under Illinois's SB 315, and the enforcement of safety obligations for general-purpose AI models under the EU AI Act — but the panel cautioned that these efforts will only matter if they are rigorously enforced rather than treated as box-ticking exercises. Panelists argued that it is essential to hold companies to their own words: to keep their stated thresholds and red lines binding, and to remind them of the commitments they have already made — above all, the promise to pause. The commitment, they stressed, must be honored as a genuine and unilateral red line, not one made contingent on the conduct of competitors. The panel maintained that the credibility of frontier AI safety can no longer rest on self-governance; it requires external verification and a floor of practice that holds companies under competitive pressure to safety.

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

Bibliography

Altman, Sam, and Jakub Pachocki. 2026. "Built to Benefit Everyone: Our Plan." *OpenAI*, June 8, 2026. https://openai.com/index/built-to-benefit-everyone-our-plan/.

Anthropic. 2026a. "Statement on the US Government Directive to Suspend Access to Fable 5 and Mythos 5." *Anthropic*. https://www.anthropic.com/news/fable-mythos-access.

Anthropic. 2026b. "When AI Builds Itself." *Anthropic*. Accessed June 25, 2026. https://www.anthropic.com/institute/recursive-self-improvement.

Bengio, Yoshua, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Trevor Darrell, Yuval Noah Harari, et al. 2024. "Managing Extreme AI Risks amid Rapid Progress." *Science* 384, no. 6698: 842–45.

Bhuiyan, Johana. 2025. "ChatGPT Encouraged Adam Raine's Suicidal Thoughts. His Family's Lawyer Says OpenAI Knew It Was Broken." *The Guardian*, August 29, 2025. https://www.theguardian.com/us-news/2025/aug/29/chatgpt-suicide-openai-sam-altman-adam-raine.

Bommasani, Rishi, Kevin Klyman, Shayne Longpre, Sayash Kapoor, Nestor Maslej, Betty Xiong, Daniel Zhang, and Percy Liang. 2023. "The Foundation Model Transparency Index." arXiv preprint arXiv:2310.12941. https://arxiv.org/abs/2310.12941.

Bommasani, Rishi, Kevin Klyman, Sayash Kapoor, Shayne Longpre, Betty Xiong, Nestor Maslej, and Percy Liang. 2024. *The Foundation Model Transparency Index v1.1: May 2024*. Stanford, CA: Stanford Center for Research on Foundation Models. https://crfm.stanford.edu/fmti/paper.pdf.

Carlini, Nicholas, Newton Cheng, Keane Lucas, Michael Moore, Milad Nasr, Vinay Prabhushankar, and Winnie Xiao. 2026. "Assessing Claude Mythos Preview's Cybersecurity Capabilities." *Anthropic*, April 7, 2026. https://www.anthropic.com/research/mythos-preview.

De Luce, Dan, Gordon Lubold, Kevin Collier, and Jared Perlo. 2026. "U.S. Military Is Using AI to Help Plan Iran Air Attacks, Sources Say, as Lawmakers Call for Oversight." *NBC News*, March 11, 2026. https://www.nbcnews.com/tech/tech-news/us-military-using-ai-help-plan-iran-air-attacks-sources-say-lawmakers-rcna262150.

Elias, Jennifer. 2026. "Google's AI Chatbot Allegedly Told User to Stage 'Mass Casualty Attack,' Wrongful Death Suit Claims." *CNBC*, March 4, 2026. https://www.cnbc.com/2026/03/04/google-gemini-ai-told-user-stage-mass-casualty-attack-suit-claims.html.

Hays, Kali. 2026. "Anthropic Boss Rejects Pentagon Demand to Drop AI Safeguards." *BBC News*, February 27, 2026. https://www.bbc.com/news/articles/cvg3vlzzkqeo.

Huang, Yue, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, et al. 2024. "TrustLLM: Trustworthiness in Large Language Models." arXiv preprint arXiv:2401.05561. https://arxiv.org/abs/2401.05561.

Liang, Percy, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2023. "Holistic Evaluation of Language Models." *Transactions on Machine Learning Research*. https://openreview.net/forum?id=iO4LZibEqW.

Manson, Katrina, and Emily Chang. 2026. "Anthropic CEO Doesn't Know If Claude Used in Iran School Strike." *Bloomberg Law*, June 10, 2026. https://news.bloomberglaw.com/artificial-intelligence/anthropic-ceo-doesnt-know-if-claude-used-in-iran-school-strike.

OpenAI. 2026. "Previewing GPT-5.6 Sol: A Next-Generation Model." *OpenAI*. June 26, 2026. https://openai.com/index/previewing-gpt-5-6-sol/.

Perset, Michel, and Sara Fialho Esposito. 2025. *How Are AI Developers Managing Risks? Insights from Responses to the Reporting Framework of the Hiroshima AI Process Code of Conduct*. Paris: OECD. https://www.oecd.org/content/dam/oecd/en/publications/reports/2025/09/how-are-ai-developers-managing-risks_fbaeb3ad/658c2ad6-en.pdf.

Phan, Long, and Dan Hendrycks. 2025. *CAIS AI Dashboard*. Center for AI Safety. Accessed October 3, 2025. https://dashboard.safe.ai.

SaferAI. 2025. "Risk Management Ratings — Frontier AI Companies." *SaferAI*. Accessed October 3, 2025. https://ratings.safer-ai.org/.

Stein-Perlman, Zach. 2025. *AILabWatch: Tracking AI Lab Research Outputs*. Accessed October 3, 2025. https://ailabwatch.org/.

Stein-Perlman, Zach. 2025. *AISafetyClaims.org: Tracking Safety Claims Made by AI Companies*. Accessed October 3, 2025. https://aisafetyclaims.org/.

World Economic Forum. 2026. "The Day After AGI." Session, World Economic Forum Annual Meeting 2026. Accessed June 25, 2026. https://www.weforum.org/meetings/world-economic-forum-annual-meeting-2026/sessions/the-day-after-agi/.

Zeng, Yi, Yu Yang, Andy Zhou, Jeffrey Ziwei Tan, Yuheng Tu, Yifan Mai, Kevin Klyman, Minzhou Pan, Ruoxi Jia, Dawn Song, Percy Liang, and Bo Li. 2024. "AIR-Bench 2024: A Safety Benchmark Based on Risk Categories from Regulations and Policies." arXiv preprint arXiv:2407.17436. https://arxiv.org/abs/2407.17436.

Appendix A: Grading Sheets

Each of our panellists were presented with the full contents of this appendix to inform their grading decisions.

The grading sheets are broken down by domain, and panellists were asked to provide grades for each company per domain. Within each domain is a set of indicators: a collection of facts about the companies.

Grading Sheets:

Risk Assessment
6 indicators

Current Harms
9 indicators

Safety Frameworks
4 indicators

Existential Safety Strategy
4 indicators

Governance & Accountability
4 indicators

Information Sharing and Public Messaging
10 indicators

Additional context on Chinese Regulatory System

How does it influence Chinese companies' behavior?

It is challenging to provide a fair comparison between frontier AI companies in China and those in the United States because of differing contexts. It is not obvious whether companies are more likely to abide by their own voluntary commitments (which are common in the U.S.) or draft laws and government standards that have not yet come into force (which are common in China). To enable our reviewers to draw their own conclusions, we will summarize the status of relevant Chinese laws and standards for each indicator.

In China, national and local regulations carry immediate force, as they carry legal and market-access consequences. Voluntary standards, while not legally binding, often serve as practical compliance references and are widely adopted in practice. Even draft regulations and policy guidance—at both national and local levels—may shape expectations and signal future directions, prompting companies to align early in order to sustain legitimacy and regulatory goodwill. In this context, the relative scarcity of voluntary safety commitments by Chinese companies may at least in part reflect differences in regulatory expectations and channels for policy engagement.

Below is a high-level summary of how each type of legislation or policy documents influence Chinese AI companies' behaviors.

National Binding Instruments

What it includes. Foundational statutes passed by the National People's Congress and its Standing Committee, including the **Cybersecurity Law (2017, amended October 2025)**, **Data Security Law (2021)**, and **Personal Information Protection Law (2021)** — plus State Council and department rules: **Provisions on the Administration of Algorithmic Recommendations in Internet Information Services (2022)**, **Provisions on the Administration of Deep Synthesis of Internet-based Information Services, Generative Artificial Intelligence Interim Measures (2023)**, **Interim Measures for the Management of Generative Artificial Intelligence Services (2023)**, **Regulations on the Management of Network Data Security (2025), Measures for Labelling of Artificial Intelligence-Generated Synthetic Content (2025)**, **Interim Measures for the Administration of Anthropomorphic Artificial Intelligence Interaction Services (2026)**, among other sectoral rules in finance, healthcare, and autonomous driving.

What it means for companies. National binding instruments are most determinative in driving company behaviors. Statutes generate governance infrastructure and fundamental principles, while departmental rules operationalize concrete implementation measures, including algorithm filings, pre-deployment security assessments, content labeling, training-data provenance logs, and requirements on refusal behaviors.

Enforcement. The Cyberspace Administration of China (CAC), the Ministry of Industry and Information Technology (MIIT), the Ministry of Public Security (MPS), the State Administration for Market Regulation (SAMR), the National Radio and Television Administration (NRTA), and sectoral regulators (e.g. People's Bank of China, National Financial Regulatory Administration, National Health Commission, National Medical Products Administration, Ministry of Transport).

Service providers face a penalty structure spanning administrative sanctions (e.g. fines up to 50 million RMB or 5 percent of turnover, service suspension, app-store delisting, license revocation), civil liability (e.g. user lawsuits and public-interest litigation), criminal liability (e.g. imprisonment and fines).

Local Binding Instruments

What it includes. Local artificial intelligence regulations enacted by provincial or municipal People's Congresses and implementing rules issued by local governments, including the **Shanghai Regulations on Promoting the Development of the Artificial Intelligence**

Industry (2022), **Shenzhen Special Economic Zone Regulations on Promoting Artificial Intelligence Industry** (2022), **Beijing Measures on Promoting Innovative Development of Artificial Intelligence** (2023), and comparable rules in Hangzhou, Chengdu, and Guangzhou.

What it means for companies. Legally these are predominantly promotional (促进) instruments and are, subordinate to national law — where they conflict, national statutes and departmental rules prevail. They create few new prohibitions; their binding bite is limited to localized duties such as ethics review of high-risk applications, risk classification, and sandbox conditions.

Enforcement. Enforcement is led by local counterparts of the national agencies (local CAC, Economy and Information Technology commissions, Public Security and Market Regulation bureaus) and bodies created by the rules themselves (AI working groups, ethics committees, sandbox administrators). Consequences are mainly incentive-based rather than punitive: companies risk losing subsidies, tax benefits, compute vouchers, and dataset access, or being expelled from sandboxes and pilots.

Voluntary Technical Standard

What it includes. Recommended national standards (GB/T), technical documents, and practice guidelines issued chiefly by the National Information Security Standardization Technical Committee (TC260) under SAMR's national standardization administration and the guidance of the CAC. Examples include the *Basic Security Requirements for Generative AI Services* (TC260-003, later GB/T 45654-2025), the *Security Specification for Generative AI Pre-training and Fine-tuning Data* (GB/T 45652-2025), the *Generative AI Data Annotation Security Specification* (GB/T 45674-2025), the *Assessment Specification for Security of Machine Learning Algorithms* (GB/T 42888-2023).

What it means for companies. Legally these standards are recommended (推荐性) and non-binding on their own. Their practical force comes from operationalizing binding regulations: GB/T 45654-2025 supplies the concrete security-assessment methodology providers use to satisfy the filing and assessment duties under the Generative AI Interim Measures, making it a de facto compliance benchmark and reference for regulators and third-party evaluators. [Zou & Zhang, 2025]

Enforcement. As recommended standards they carry no standalone penalties; TC260 and the standardization administration are standard-setters, not enforcers. In practice, conformity is effectively compelled because regulators reference these standards when judging compliance with binding regimes. Therefore, failure to meet a standard surfaces as a failure under the underlying regulation, such as the Interim Measures, rather than as a breach of the standard itself. Consequences therefore flow back through the national binding instruments.

Draft Regulations and Standards

What it includes. Instruments released for public consultation but not yet in force, including draft departmental rules and measures and draft national/industry/municipal standards from TC260 and other bodies (e.g. draft safety, evaluation, and labeling standards before finalization).

What it means for companies. Legally non-binding while in draft — they create no enforceable obligations and cannot be the basis for sanctions. Their significance is anticipatory: they signal the direction and likely content of forthcoming binding rules, give a window to submit comments and shape final text, and let companies pre-position compliance.

Enforcement. Not enforced. There are no penalties for non-conformance with a draft. The relevant bodies (CAC, MIIT, TC260, sometimes the NPC for the AI Law) collect feedback and revise.

Strategic and Policy Guidance Documents

What it includes. High-level, non-binding documents setting national direction, including the State Council's New Generation Artificial Intelligence Development Plan (2017), the "AI+" initiative guidance, MIIT and NDRC industrial-policy notices, ethical frameworks such as the New Generation AI Ethics Specifications (2021) and the Ethical Norms for New Generation AI, and foreign-facing positions like the Global AI Governance Initiative (2023). These also include high-profile political speeches or party directives from senior leadership.

What it means for companies. No direct legal obligations or liability attached to these documents. Their relevance is strategic: this high-level policy guidance functions as a behavioral steering tool, compelling platform firms to anticipate regulatory trends, publicly align with state priorities, and adjust business practices long before formal laws are enacted.

Enforcement. Not legally enforced. Influence is exercised through the policy and fiscal levers of issuing bodies (State Council, NDRC, MIIT, MOST, CAC, TC260, and the science-and-tech ethics committees): alignment shapes eligibility for subsidies, tax incentives, pilots, and procurement, while divergence risks lost support and reputational/political friction rather than formal sanction.

Domain

Risk Assessment

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

Table of Contents

At present, no binding national regulations or standards—whether mandatory or recommended—explicitly address frontier AI risks or define corresponding risk assessment processes. The Shanghai Draft offers early compliance guidance but its final scope, adoption timeline if adopted at the national level, and extrajurisdictional applicability remain uncertain. Nonetheless, the AI Safety Governance Framework 2.0 signals the government's intent to establish national standards to systematically address frontier risks in the near future.

National binding instruments and voluntary technical standards are not applicable here.

Local Binding Instruments

Shenzhen Regulation (2022) requires high-risk AI applications to adopt a regulatory model of ex-ante assessment and risk warning (Article 66), although it doesn't specify which risks the service providers should assess. This does not apply to Z.ai's GLM models (Beijing) or Deepseek's R1 model (Zhejiang), and Alibaba's Qwen models (Zhejiang).

Draft Regulations and Standards

Article 5.8 of the Shanghai draft recommended standard on multi-modal LLMs safety evaluations enumerates potential high-risk capabilities of large models, including generation of malicious software, enabling the development of biological or chemical weapons, engaging in deceptive behavior, and exhibiting self-replication or self-improvement tendencies.

However, Article 6.1.8 narrows the focus to cyber-related risks, requiring evaluation of the model's potential to uplift cyberattacks—specifically through the generation of malicious code, phishing emails, password cracking, vulnerability exploitation, and social-engineering attacks.

Article 7 of the Draft covers three main aspects: evaluation methods, evaluation procedures, and reporting requirements. For methods, it outlines distinct evaluation approaches for text, image, voice, and video generation. For procedures, it specifies four key steps: establishing an evaluation committee, determining the scope and content of evaluation, conducting the evaluation work, and producing the final evaluation report. For reporting, it requires detailed documentation of methodologies (including automated testing, manual review, and user feedback mechanisms), analysis of false negatives and false positives, and concrete improvement suggestions. The final report must include both quantitative data and illustrative materials such as diagrams and case studies.

The recommended national standard on model alignment is currently under review: **Artificial intelligence—Large language model alignment capability evaluation.**

Strategic and Policy Guidance Documents

The AI Safety Governance Framework 2.0

Article 5.8 calls for the establishment of an AI safety evaluation system that integrates model and algorithm safety testing, general application safety testing, and scenario-specific safety testing.

Article 3.2.3 (c) explicitly calls for focusing on risk including loss of control over knowledge and capabilities of nuclear, biological, chemical, and missile weapons.

Specifically, Article 6.1.9 recommends regular safety evaluations and testing where a risk classification, grading, and optimization mechanism is established, clearly defining testing objectives, scope, and safety dimensions before each evaluation. It calls for the development of diverse testing datasets that cover a wide range of application scenarios, and the formulation of targeted model optimization strategies for different categories of risks.

Moreover, Article 5.11 calls for building global consensus and coordination mechanisms to address AI loss-of-control risks. It emphasizes strengthening end-use management of AI systems by setting specific safeguards for their application in nuclear, biological, chemical, and missile-related domains to prevent misuse. The clause promotes the adoption of trusted AI principles that integrate technical, ethical, and managerial dimensions, aiming to foster broad international alignment on responsible AI governance. It also requires developers to conduct regular testing to assess whether their models may pose potential technical loss-of-control risks.

Internal

Indicator

Dangerous Capability Evaluations

Definition

This indicator assesses whether organizations conduct systematic evaluations of dangerous capabilities before deploying frontier models. Priority domains include biological and chemical weapons, offensive cyber operations, recursive self-improvement risks, and behaviors associated with goal misalignment or deception. Evidence is drawn from model cards detailing testing methodologies and results. The focus is on external deployments, as there is insufficient transparency on internal deployments

Why it matters

Systematic evaluations for high-risk capabilities reflect institutional responsibility for managing low-probability, high-impact harms. In contrast to more routine risks—where market forces often suffice—frontier threats require deliberate foresight. Firms that fail to test for these dangers risk contributing to unmanaged systemic vulnerability.

Relevant Regulations and Standards

EU AI Code of Practice (Measure 3.2 + Appendix 3.1)

Signatories will conduct at least state-of-the-art model evaluations in the modalities relevant to the systemic risk to assess the model's capabilities, propensities, affordances, and/or effects.

Model evaluations should be designed and conducted using methods that are appropriate for the model and the systemic risk and should include open-ended testing of the model.

Examples of model evaluation methods include: Q&A sets, task-based evaluations, benchmarks, red-teaming and other methods of adversarial testing, human uplift studies, model organisms, simulations, and/or proxy evaluations for classified materials.

The evaluation should ensure 1) internal validity, 2) external validity, 3) reproducibility.

California SB 53 § 22757.12(c)(2)(A)–(D); New York RAISE Act § 1421(3)(b); Illinois SB 315 § 10(c)(2)(A)–(D)

Before, or concurrently with, deploying a new frontier model or a substantially modified version of an existing frontier model, a large frontier developer shall include in the transparency report summaries: (A) Assessments of catastrophic risks from the frontier model conducted pursuant to the large frontier developer's frontier AI framework. (B) The results of those assessments. (C) The extent to which third-party evaluators were involved. (D) Other steps taken to fulfill the requirements of the frontier AI framework with respect to the frontier model.

Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud	Mistral
Claude Opus 4.7	GPT 5.5	Gemini 3.1 Pro *	Muse Spark	Grok 4.1	V4	GLM 5	Qwen 3.5 Max	Mistral Large 3
Biosecurity & Chemical Risk								
The company has released the same model version that the final round of safety (framework) evaluations were conducted on. **Safety Framework Decision** The evaluations conclude that (1) It is hard to be confident regarding whether a model passes CB1 (non-novel CB weapons production capabilities) thresholds; (2) Capabilities evaluations demonstrate that Opus 4.7 can provide information relevant to the threat model, is capable of significant cross-domain synthesis relevant to catastrophic biological weapons development, therefore requiring mitigation measures equal to or stronger than ASL-3 protections. (2) Opus 4.7 does not surpass CB2 (novel CB weapons production capabilities) thresholds. **Rationale** (1) Expert red-teamers found "significant strengths in the synthesis of the published record" but "weakness in the model's utility in endeavors requiring novel approaches" — "lack of anticipatory behavior," "insufficient depth in protocol development," and "overconfidence in the feasibility of synthesis steps." Catastrophic scenarios required "significant guidance…to steer the model." (2) On sequence-to-function, Opus 4.7 "trailed Claude Mythos Preview on both tasks" **The Evaluation Scope covers:** (1) Expert red-teamings (Known CB weapons, n=9 biology + 2 chemistry experts); (2) Long-form virology tasks (Known biological weapons); (3) Multimodal virology (VCT) (Known biological weapons); (4) DNA Synthesis Screening Evasion (Known biological weapons); (5) Sequence-to- function modeling and design (Partnership with Dyno Therapeutics, Novel biological weapons) **Methodological Details** include: (1) Environment and elicitation (e.g. tools and agentic harnesses, extended thinking); (2) Model selection (multiple snapshots, including "helpful-only" variants (harmlessness removed)); (3) Evaluation designs (e.g. uplift and feasibility scoring rubrics used in expert red-teaming); (4) Quantitative benchmarks (compared to expert scoring, as well as other applicable models); (5) Design rationale (e.g. focus on multi-step, medium-timeframe scenarios). (Source: Opus 4.7 System Card pp.15-25)	*The company has released the same model version that the final round of safety (framework) evaluations were conducted on.* **Safety Framework Decision** GPT-5.5 is treated as High capability and below Critical. Bio capabilities are prioritized given the higher potential severity of biological threats. **Rationale** The model falls under pre-determined quantitative thresholds on the autonomous wet-lab/design proxies most relevant to Critical capability but exceeds baselines on some knowledge-oriented evaluations. The **Evaluation Scope** includes (1) Multimodal Troubleshooting Virology (350 SecureBio held-out questions); (2) ProtocolQA Open-Ended (108 short-answer items, 19-PhD baseline); (3) Tacit Knowledge & Troubleshooting MCQ (Gryphon Scientific, uncontaminated, all 5 biothreat-creation stages); (4) TroubleshootingBench (52 expert-transcribed protocols × 3 questions, 12-PhD baseline, 80th-percentile expert = 36.4%); (5) Biochemistry knowledge delta vs. GPT-5.4-thinking; (6) Hard-negative protein binder discrimination (43 targets, 492 hotspots, ipTM ≥ 0.8); (7) DNA sequence design for transcription factor binding (11 TFs from Nucleobench, vs. Ledidi); (8) External evaluations from SecureBio and US CAISI; (9) Bio Bug Bounty Program for post-deployment universal-jailbreak elicitation. **Methodological Details** include: (1) Environment & elicitation: e.g. agentic harnesses with computer/browser access for protein binder tasks; (2) Model selection with multiple pre-release checkpoints; (3) Evaluation design; (4) Quantitative benchmarks: e.g. comparisons against PhD expert consensus baselines; (5) Design rationale: e.g. safeguard testing with rubrics mapped to OpenAI's bio-risk taxonomy. (Source: GPT 5.5 System Card pp.21-28)	*The company further modified the model after the final round of safety (framework) evaluations but explicitly mentioned and described all further changes in the model documentation.* **Safety Framework Decision** The model (Deep Think Mode) is below the alert threshold and below the CBRN Uplift Level 1 CCL. **Rationale** The model can provide highly accurate and actionable information but still fails to offer novel or sufficiently complete and detailed instructions for critical stages, to significantly enhance the capabilities of low to medium resourced threat actors required for the CCL. The **Evaluation Scope** includes: (1) Knowledge-based and reasoning Multiple Choice Questions (MCQs) using external benchmarks including LAB-Bench (with Cloning Scenarios) and SecureBio VCT for virology troubleshooting; (2) Single and static multi-turn Open-Ended Questions (OEQs) tailored across CBRN domains and threat actor profiles; (3) Expert-designed red-teaming scenarios broken down by attack stage, benchmarked against a web baseline; (4) External "wet lab" uplift trial preliminary results from Panoplia Laboratories; (5) External Safety Testing by third-party evaluators across CBRN domains. **Methodological Details** include: (1) Environment & elicitation: e.g. multi-turn OEQs spanning varied threat actor profiles, resource levels, equipment, and harm intent; (2) Model selection: Deep Think Mode; (3) Evaluation design: e.g. holistic, expert-led decision-making; (4) Design rationale: e.g. wet-lab trials used to address the "open question on the disconnect between model evaluations and real-world risk" given that tacit laboratory knowledge is poorly captured by text-based evals. (Source: Gemini 3.1 Pro System Card pp.8, Gemini 3 Pro FSF Report, pp.5-8)	*The company has released the same model version that the final round of safety (framework) evaluations were conducted on.* Safety Framework Decision Unmitigated Muse Spark meets the "high risk" threshold for Chem & Bio risks, but with implemented mitigations the residual risk is reduced to "moderate or lower," meeting the threshold for responsible deployment. **Rationale** Across the aggregate set of dual-use and high-risk capability evaluations, Muse Spark's performance is sufficiently high that Meta cannot rule out the possibility that deployment without appropriate mitigations would substantially contribute to one or more of the identified Threat Scenarios. Specific areas of concern include performance on sophisticated dual-use evaluations and direct evidence of high-risk capabilities that could enable threat actors to acquire, produce, or release harmful chemical or biological agents. The **Evaluation Scope** includes: (1) Dual-Use Scientific Knowledge and Wet-Lab Debugging (2) Dual-Use: Agentic Scientific Capabilities (3) Dual-Use: Overall Interpretation **Methodological Details** include: (1) Environment & elicitation: e.g. isolated compute cluster for sensitive evaluations with no external tools (lower-bound estimate); (2) Model selection: Muse Spark, Muse Spark HO (helpful-only), Meta AI; (3) Evaluation design with quantitative benchmarks; (4) Quantitative comparison with a reference class of models. (Source: pp.15-33)	**Safety Framework Decision** The system card does not mention an assigned risk tier for its model in compliance with the FMF. It presents evaluations along the three directions defined by the FMF: abuse potential, concerning propensities, and dual-use capabilities. Bio-related capabilities are emphasized because they have likely the potential for the greatest scale of harm and frontier models may significantly lower the entry barrier. Grok 4.1 achieves broadly similar results to Grok 4 (and others). The **Evaluation Scope** includes: (1) General knowledge; (2) Troubleshooting incorrect laboratory protocols and failed experiments; (3) Understanding scientific papers; (4) Genetic cloning, and (5) Dual-use chemical knowledge. **Methodological Details** include: (1) Dataset scoping: e.g. text-only questions for WMDP and VCT; (2) Quantitative metrics and human baselines; (3) Quantitative comparison with other models. (Source: pp.3-4)	**Not Mentioned**	The company has released the same model version that the final round of safety (framework) evaluations were conducted on. **Evaluation details are not mentioned**	**Not Mentioned**	**Not Mentioned**

* Additional information from Gemini 3 Pro Frontier Safety Framework Report

Table continues on next page

Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud	Mistral
Claude Opus 4.7	GPT 5.5	Gemini 3.1 Pro *	Muse Spark	Grok 4.1	V4	GLM 5	Qwen 3.5 Max	Mistral Large 3
Cybersecurity Risk								
RSP Decisions not applicable. Overall cyber risk assessment reports the risk to be "Claude Opus 4.7 is roughly similar to Opus 4.6 in cyber capabilities." It also deliberate training-time suppression: "during training [Anthropic] experimented with efforts to differentially reduce these capabilities. **Acknowledged limitations** The latest frontier models have saturated nearly all of the CTF-style evaluations already, prompting exploration of new metrics. The **Evaluation Scope** covers: (1) Frontier red teaming a. Cybench (35-challenge CTF subset); b. CyberGym (1,507-task targeted vulnerability reproduction); c. Firefox 147 shell exploitation (50 crash categories, 5 trials each); (2) External testing from UK AISI (corporate network attack simulation). **Methodological Details** include: (1) Environment & elicitation: no thinking, default effort/temperature/top_p; "think" tool for interleaved multi-turn reasoning; updated harness parameters (2) Evaluation design: pass@1 aggregate (CyberGym); pass@10 (Cybench); three-level grading 0/0.5/1.0 for partial vs. full exploit (Firefox); human-hour-equivalent estimates (UK AISI) (3) Quantitative benchmarks: Cybench pass@1 96%; CyberGym ≈Opus 4.6; Firefox partial-control >2× Opus 4.6 but well below Mythos; UK AISI ~5h of 10h+ range (4) Design rationale: maintain Mythos eval suite for comparability; flag CTF saturation as motivation for new metrics. (Source: Opus 4.7 System Card pp.48-52)	**Safety Framework Decision** GPT-5.5 is treated as High capability and below Critical. **Rationale** GPT-5.5 shows meaningful gains in end-to-end exploitation and long-horizon vulnerability research, but falls short of Critical because it could not independently produce functional full-chain exploits against hardened real-world targets . The **Evaluation Scope** includes (1) Professional-level CTF) challenges across web/rev/pwn/crypto/misc, pass@12 over 16 rollouts (saturated by GPT-5.5); (2) CVE-Bench v1.0 (34/40 challenges, zero-day prompt configuration, no source-code access, pass@1 over 3 rollouts); (3) Cyber Range (15 emulated end-to-end network scenarios, pass/fail over 16 trials); (4) VulnLMP (open-ended frontier eval against widely deployed hardened software including browser targets, with high test-time-compute and verifier-owned evidence channels); (5) External evaluations from Irregular (atomic challenge suite + CyScenarioBench), US CAISI, and UK AISI (narrow cyber tasks + cyber ranges with 50M / 100M token limits) . **Methodological Details** include: (1) Environment & elicitation: e.g. agentic harness with headless Linux box and preinstalled offensive tools; (2) Model selection: representative launch checkpoint plus a reduced-refusals checkpoint; (3) Evaluation design; (4) Quantitative benchmarks: comparisons against multiple other models and external baselines; (5) Design rationale: e.g. layered safeguard stack. (Source: GPT 5.5 System Card, pp. 28-34)	**Safety Framework Decision** Gemini 3.1 Pro reaches the alert threshold but does not reach the Cyber Uplift Level 1 CCL, with Deep Think Mode performing significantly worse compared to the model without the mode. **Rationale** Additional testing on 3.1 Pro shows an increase in cyber capabilities compared to 3.0 Pro, but does not amount to CCL. The **Evaluation Scope** includes: (1) The "key skills" benchmark v1 — 50 challenges (12 hard) developed with a third party, mapped to four key skill areas: Reconnaissance, Tool Development, Tool Usage, and Operational Security; (2) The "key skills" benchmark v2 — a new set of harder, more realistic end-to-end challenges simulating full multi-stage attacks, expanding coverage to all seven key attack kill chains from Rodriguez et al. 2025; (3) External Safety Testing by third-party evaluators on operationally relevant scenarios loosely aligned with MITRE ATT&CK and the Unified Kill Chain taxonomies. **Methodological Details** include: (1) Environment & elicitation: e.g. agent harness prompting structured planning and tool use, with Bash/PowerShell command execution, Python scripts, and web search; (2) Model selection: Gemini 3.1 Pro evaluated both with and without Deep Think mode; (3) Evaluation design (4) Quantitative comparisons between various Google Gemini models; (5) Design rationale: e.g. the benchmark is grounded in the cyber evaluation framework of Rodriguez et al. 2025, drawing on real-world threat intelligence to identify representative attacks. (Source: Gemini 3.1 Pro System Card pp.8, Gemini 3 Pro FSF Report, pp.8-10)	**Safety Framework Decision** Overall risk level is moderate. **Rationale** Muse Spark does not materially alter the cyber threat landscape for end-to-end network compromise, scaled exploitation of critical vulnerabilities, or scaled Frauds and Scams. The **Evaluation Scope** includes: (1) Knowledge-based capabilities (2) Agentic cyber capabilities, including CTF challenges, complex real-world challenges, and social engineering capabilities; (3) Refusals on capability evaluations and high-severity outcome enabling prompts; (7) Insecure code propensity evaluation. **Methodological Details** include: (1) Environment & elicitation: e.g. ReAct agent with Bash and Python tools, default Kali Linux environment with optional installation; (2) Model selection: Muse Spark, Meta AI Instant, Meta AI Thinking (3) Evaluation design with quantitative benchmarks; (4) Quantitative comparison with a reference class of models. (Source: pp.34-49)	**Safety Framework Decision** The system card does not mention an assigned risk tier for its model in compliance with the FMF. It presents evaluations along the three directions defined by the FMF: abuse potential, concerning propensities, and dual-use capabilities. Grok 4.1 performs similarly to other frontier models, but substantially below the level of human cybersecurity experiments. The **Evaluation Scope** includes: (1) WMDP Cyber; (2) CyBench. **Methodological Details** include: (1) Quantitative comparison with other models. (Source: pp.3-4)	**Not Mentioned**	GLM-5 scores 43.2% on CyberGym, referenced as "a public benchmark that evaluates whether AI can find vulnerabilities in real open-source software." However, this is not related to risk assessment. (Source: GLM-5 Technical Report pp.23)	**Not Mentioned**	**Not Mentioned**

* Additional information from Gemini 3 Pro Frontier Safety Framework Report

Table continues on next page

Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud	Mistral
Claude Opus 4.7	GPT 5.5	Gemini 3.1 Pro *	Muse Spark	Grok 4.1	V4	GLM 5	Qwen 3.5 Max	Mistral Large 3
Autonomous AI R&D								
RSP Decision The evaluation concludes that the model "does not cross the Automated R&D capability threshold." (1) Autonomy threat model 1 (early-stage misalignment risk) is applicable. (2) Autonomy threat model 2 (risks from automated AI R&D) is not applicable. **Rationale** (1) No sustained, 2× speedup observed in capabilities over time (measured on a fork by Epoch Capability Index); and (2) It does not seem close to being able to fully substitute for Research Scientists and Research Engineers, especially relatively senior ones. (*Note*: Qualitative judgment by RSO based on interactions with employees and observations of research workflows and progress) **The Evaluation Scope covers:** (1) AECI capability-trajectory tracking; (2) Qualitative RSO judgment from day-to-day internal usage; (3) Supporting/Contextual: a. Task-based capability evaluations - not used for RSP determination; b. Internal productivity surveys — Mythos uplift poll + L4-substitution survey; c. Failure-mode case studies — manual reports + automated transcript scan; d. Reward hacking transcript audits. **Methodological details include:** (1) Environment & elicitation: standard + experimental scaffolds; multi-turn agentic transcripts (2) Evaluation design (e.g. unbounded scoring, AECI slope-ratio etc.) (3) Quantitative benchmarks: Kernel, Time-Series, LLM-training, Quadruped RL, Novel Compiler, Internal Suite 2; (4) Design rationale: isolate AI-attributable acceleration, not aggregate lab pace. (Source: Opus 4.7 System Card pp.25-43)	**Safety Framework Decision** GPT-5.5 is treated as below High capability in AI Self-Improvement and "does not have a plausible chance of reaching a High threshold." **Rationale** GPT-5.5 shows only modest improvements over GPT-5.4 Thinking on internal research and engineering benchmarks. The **Evaluation Scope** includes: (1) Monorepo-Bench (PR-style contributions in a large internal repository, hidden unit-test grading); (2) MLE-Bench (30 curated Kaggle competitions <50GB / <10h, bronze-medal threshold for success); (3) Internal Research Debugging Evaluation (41 real bugs from internal OpenAI research experiments + 6 alignment-auditing tasks for rediscovering misaligned behaviors); (4) OPQA / OpenAI-Proof Q&A (20 internal research and engineering bottlenecks each representing ≥1 day of senior-staff delay, pass@1 with code/log/artifact container access). **Methodological Details** include: (1) Environment & elicitation: e.g. agentic rollouts with command-line and Python tool access; container access with code and run artifacts for OPQA ; (2) Evaluation design: e.g. time-horizon decomposition for the debugging eval; (3) Quantitative comparisons with other GPT models; (4) Design rationale: e.g. framing capability evals as a lower bound. (Source: GPT 5.5 System Card, pp. 34-37)	**Safety Framework Decision** Gemini 3.1 Pro (Deep Think mode) is below the alert threshold and below both the ML R&D Acceleration Level 1 and Automation Level 1 CCLs. **Rationale** The model shows gains on RE-Bench compared to Gemini 3 Pro. However, its average performance across all challenges remains beneath the alert threshold for the CCLs. The **Evaluation Scope** includes: (1) The Research Engineering Benchmark (RE-Bench) from Wijk et al. 2024 — a collection of 7 ML research tasks each requiring several hours of human engineering effort, used as a leading indicator for both Automation Level 1 and Acceleration Level 1 CCLs; (2) Two RE-Bench challenges (Finetune GPT-2 for QA and Scaffolding for Rust Codecontest) are omitted because they require internet access; (3) External Safety Testing by third-party evaluators assessing potential for catastrophic harm via sabotaging AI research, focused on a scheming AI's ability to subvert successor-system development during autonomous ML R&D deployment. **Methodological Details** include: (1) Environment & elicitation: e.g. METR "Modular" scaffold with minimal modifications (2) Model selection: Deep Think mode; (3) Evaluation design (4) Quantitative comparisons across various Gemini models, with scores normalised against human-written solutions; (5) Design rationale: e.g. RE-Bench low performance (narrow set of tasks) is treated as evidence the model is unlikely to meaningfully accelerate AI progress. (Source: Gemini 3.1 Pro System Card pp.8, Gemini 3 Pro FSF Report, pp.13-15)	**Safety Framework Decision** Meta assesses Loss of Control risk as "moderate or lower," including pre-deployment evaluation reliability, automated monitoring reliability, and misalignment-related propensities. **Rationale** Muse Spark's autonomous technical capabilities — spanning software engineering, ML engineering, and end-to-end research — remain too constrained for the kind of sustained, creative, long-horizon work that could accelerate AI development beyond Meta's ability to evaluate it. The **Evaluation Scope** includes: (1) Pre-deployment assessment reliability, considering accelaration for AI development and selective model performance in catastrophic risk domains; (2) Reliable monitorability; (3) Misaglined propensity; (4) Emerging outcomes, including AI R&D sabotage and instrumental self-exfiltration. **Methodological Details** include: (1) Environment & elicitation: e.g. ReAct scaffolds with bash and ML-specific tools for AI R&D benchmarks, (2) Evaluation design with quantitative benchmarks; (3) Quantitative comparison with a reference class of models; (4) Design rationale: e.g. data contamination mitigated for MLE-Bench by excluding competitions from training data and preventing internet access. (Source: pp.50-70)	Not Mentioned	Not Mentioned	Not Mentioned	Not Mentioned	Not Mentioned

* Additional information from Gemini 3 Pro Frontier Safety Framework Report

Table continues on next page

Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud	Mistral
Claude Opus 4.7	GPT 5.5	Gemini 3.1 Pro *	Muse Spark	Grok 4.1	V4	GLM 5	Qwen 3.5 Max	Mistral Large 3
Scheming & Misalignment Risks								
Safety Framework Decision Overall alignment risk assessment reports the risk to be "very low, but higher than for models prior to Claude Mythos Preview." **Rationale** (1) Opus 4.7 "appears to have similar overall alignment properties to Claude Opus 4.6" and is "less capable than Claude Mythos Preview." (2) Covert capabilities are "slightly improved over Claude Opus 4.6 but are still roughly comparable." (3) General-access deployment adds two external pathways, but the 4.6→4.7 capability jump is insufficient to raise risk meaningfully. The **Evaluation Scope** covers: (1) Background expectations (deployment/ affordances; prior-model continuity; sources of misalignment); (2) Training-time monitoring (RL/SFT episode review; environment evaluation; CoT-supervision bug, 7.8%) (3) Covert capability evaluations (SHADE-Arena; Minimal-LinuxBench; rare-action; secret-password); (4) Behavioral alignment evaluations (alignment audits; realism-steering deception check; pilot internal-use review); (5) Deployment-time monitoring (internal usage monitoring system; coverage/recall improvements since 4.6); (6) Risk-pathway analysis (six pathways inherited from Mythos + Pathway 7: other AI developers' R&D; Pathway 8: major governments) **Methodological details** include: (1) Environment & elicitation (e.g. extended-thinking and CoT-visible monitoring conditions) (2) Evaluation design (3) Quantitative benchmarks (4) Design rationale (Source: Opus 4.7 System Card pp.	**Safety Framework Decision** Misalignment-related capabilities (sandbagging) fall under the Preparedness Framework's Research Categories. The system card concludes that "[OpenAI does] not currently have evidence that GPT-5.5 has misalignment propensities or the long-range autonomy needed to cause internal deployment risks such as successfully self-exfiltrating or sabotaging internal research." **Rationale** Internal evals indicate GPT-5.5 is "slightly more misaligned" than previous models but "no evidence of high severity"; Apollo Research similarly "did not find evidence that the evaluated checkpoint poses substantially elevated risk of catastrophic harm via scheming relative to tested baseline models." The **Evaluation Scope** includes: (1) Deception evaluations on prompts representative of external ChatGPT usage; (2) Evaluation of Coding Agents for Misalignment by Resampling Internal Traffic; (3) Chain-of-Thought evaluations on monitorability and controllability; (4) External eval by Apollo Research on strategic deception, in-context scheming, sandbagging, and sabotage. **Methodological Details** include: (1) Environment & elicitation: e.g. tool-response simulator validated against original trajectories; (2) Model selection: pre-release checkpoints; (3) Evaluation design: e.g. severity scoring, subcategory grouping; (4) Quantitative comparisons on internal evaluations; (5) Design rationale: resampling framed as an indicator of internal-deployment risk; CoT controllability tracked as an "early warning sign" of models' ability to obscure reasoning. (Source: GPT 5.5 System Card, pp.12-19, 38)	**Safety Framework Decision** Gemini 3.1 Pro (Deep Think mode) is below the alert threshold and below the Instrumental Reasoning Level 1 and Level 2 CCLs. (Misalignment remains an exploratory risk area, and CCLs in this domain are not associated with explicit risk acceptance criteria given the nascent nature of the research.) **Rationale** On stealth evaluations, Gemini 3.1 Pro performs similarly to Gemini 3 Pro. On situational awareness, the model is stronger than Gemini 3 Pro, although performance is inconsistent. The **Evaluation Scope** includes: (1) 4 stealth challenges from Phuong et al. 2025; (2) 11 situational awareness challenges from Phuong et al. 2025, (3) CoT legibility analysis using metrics from Emmons et al. 2025; (4) Correctness checks including automated sandbagging detection across all domains, plus manual transcript inspection in ML R&D and misalignment for evaluation awareness; (5) External Safety Testing by third-party evaluators on autonomous systems risks, focusing on strategic deception and red-teaming scenarios mimicking internal AI R&D for sabotage propensity. **Methodological Details** include: (1) Environment & elicitation: e.g. scaffolded models using chain-of-thought reasoning and reflection to enhance context utilisation and planning (2) Model selection: Deep Think mode; (3) Evaluation design: e.g. normalised scoring on stealth and situational awareness suites (4) Quantitative comparisons across various Gemini models, with scores normalised against human-written solutions. (Source: Gemini 3.1 Pro System Card pp.8, Gemini 3 Pro FSF Report, pp.13-15)	**Included in the "Autonomous AI R&D" section above.**	**Safety Framework Decision** Misalignment-adjacent risks are evaluated under the RMF's "concerning propensities" category, focused specifically on deception and manipulation. No formal threshold or tier is assigned. Grok 4.1 increases in both manipulation and deception. The **Evaluation Scope** includes: (1) MASK; (2) Sycophancy. **Methodological Details** include: (1) Quantitative comparison with other models. (Source: pp.2-3)	**Not Mentioned**	**Not Mentioned**	**Not Mentioned**	**Not Mentioned**

* Additional information from Gemini 3 Pro Frontier Safety Framework Report

Table continues on next page

Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud	Mistral
Claude Opus 4.7	GPT 5.5	Gemini 3.1 Pro *	Muse Spark	Grok 4.1	V4	GLM 5	Qwen 3.5 Max	Mistral Large 3
Manipulation Risks								
Safety Framework Decision Not applicable. **Conclusion and Rationale** The model does not have the level of autonomous capability needed for full persona and network management, coordinated content delivery, and scaled social engineering campaigns. Although the model has scored higher on evals than both Opus 4.6 (both tasks) and Mythos Preview (one task), it still requires substantial human direction would be required for most operational steps. The **Evaluation Scope** includes: (1) Voter suppression scenario; (2) Domestic polarization scenario. **Methodological Details** include: (1) Environment and elicitation (e.g. 3 times with 3 different levels of simulated platform friction); (2) Helpful-only variants (3) Quantitative metrics and benchmarks. (Source: pp.80-81)	**Not Mentioned**	**Safety Framework Decision** Gemini 3.1 Pro (Deep Think mode) is below the alert threshold and below the Harmful Manipulation Level 1 (exploratory) CCL. **Rationale** Gemini 3.1 Pro showed higher manipulative efficacy for belief change metrics compared to a non-AI baseline, which is the same as Gemini 3 Pro and did not reach the alert threshold. The **Evaluation Scope** includes: (1) A propensity evaluation based on (a) fully synthetic multi-turn dialogues following Ibrahim et al. 2025, and (b) experimental logs from the efficacy evaluation, designed to detect how frequently a model engages in manipulative mechanisms outlined in El-Sayed et al. 2024; (2) An efficacy evaluation via a human behavioural study; (3) 2 experimental conditions per study — a control condition and an adversarial condition; (4) External Safety Testing by third-party evaluators on radicalisation, extremism, and democratic harms. **Methodological Details** include: (1) Environment & elicitation: e.g. multi-turn synthetic user simulations for propensity assessment (2) Model selection: Deep Think mode; (3) Evaluation design: e.g. ethics review via HuBREC (Human Behavioural Research Ethics Committee); (4) Quantitative comparisons: e.g. Gemini 3.1 Pro's maximum odds ratio of 3.6x matches Gemini 3 Pro; (5) Design rationale: e.g. the FSF focuses on outcome harms rather than process harms given its capability-focused nature. (Source: Gemini 3.1 Pro System Card 8, Gemini 3 Pro FSF Report pp.8-10)	**Not Mentioned**	**Included in the "Scheming & Misalignment Risk" section above.**	**Not Mentioned**	**Not Mentioned**	**Not Mentioned**	**Not Mentioned**

* Additional information from Gemini 3 Pro Frontier Safety Framework Report

Indicator

Elicitation for Dangerous Capability Evaluations

Definition

This indicator assesses how clearly a company explains its elicitation strategy, which is the systematic and state-of-the-art techniques it uses to reveal the model's full range of capabilities and potential dangerous behaviors that may otherwise remain concealed. Such techniques include adapting test-time compute, rate limits, scaffolding, and tools, and conducting fine-tuning and prompt engineering.

Why it matters

Standard evaluations often capture only a model's default, surface-level behavior, leaving deeper or more hazardous capabilities undiscovered. By systematically varying prompts, sampling methods, tools, and system configurations, evaluators can reveal capabilities that may emerge only under real-world or adversarial conditions. A comprehensive, transparent, and well-resourced approach demonstrates a credible commitment to risk discovery.

Relevant Regulations and Standards EU AI Code of Practice Safety and Security	**Anthropic** Claude Opus 4.7	**OpenAI** GPT 5.5	**Google DeepMind** Gemini 3.1 Pro *	**Meta** Muse Spark	**xAI** Grok 4.1	**DeepSeek** V4	**Z.ai** GLM 5	**Alibaba Cloud** Qwen 3.5 Max	**Mistral** Mistral Large 3
Appendix 3.2 Signatories are required to conduct model evaluations using at least state-of-the-art elicitation methods that minimize under-elicitation and model deception during model evaluation, and that match both the capabilities of potential misuse actors and the model's expected use context. Examples of the measures include adapting test-time compute, rate limits, scaffolding, tools, fine-tuning, and prompt engineering.	**Adapting test-time compute** is reported in CBRN evaluations (e.g. extended thinking mode used in most evaluations), cyber evaluations (e.g. pass@1 with 10 trials per challenge on Cybench). **Scaffolding with tool use** is reported in CBRN evaluations (e.g. agentic harnesses for long-form virology tasks equipped with domain-specific bio tools and search/research tools), cyber evaluations (e.g. a testing harness wrapping a SpiderMonkey shell to mimic a Firefox 147 content process, with a "think" tool available to the agent). **Helpful-only variants** are reported in CBRN evaluations and AI R&D evaluations.	**Adapting test-time compute / parallel attempts** is reported to Cybersecurity evaluations (e.g. pass@12 over 16 rollouts for CTF challenges), Bio/Chem evaluations (e.g. pass@4 for hard-negative protein binding), and AI R&D (e.g. multi-rollout steps). **Scaffolding with tool use** are reported in Cybersecurity evaluations (e.g. headless Linux boxes with preinstalled offensive security tools, run via command-execution harnesses on CTF-style challenges), Bio/Chem evaluations (e.g. agents with computer and browser access operating on PDB file inputs for protein binding tasks), and AI R&D evaluations (e.g. agents operating in virtual environments with GPU access). **Helpful-only / reduced-refusal model variants** are reported in Cybersecurity evals (reduced-refusal checkpoints provided to external researchers) and Bio/Chem (reduced-refusal checkpoints provided to external researchers with API content filtering disabled) **Open-ended / multi-turn qualitative probing** is reported in Bio/Chem and Cybersecurity in external researchers' description.	The Gemini 3.1 Pro system card has mentioned: **Adapting test-time compute** across all risk domains via Deep Think Mode. Since the system card has explicitly mentioned that "more details on evaluations ... can be found in the Gemini 3 Pro Frontier Safety Framework Report," we have also included the elicitation strategies from the FSF report: **Scaffolding with tool use** in cyber evaluations (e.g. a custom controller enforcing a multi-stage thinking protocol paired with Bash + Python on Linux, PowerShell on Windows), ML R&D evaluations (e.g. METR's "Modular" scaffold from Wijk et al. 2024 applied with minimal modifications and implicit code-execution tooling), and misalignment evaluations (e.g. scaffolds equipped with chain-of-thought reasoning and reflection). **Iterative prompting / multi-turn elicitation** is reported in CBRN evaluations (e.g. single and static multi-turn open-ended questions tailored to threat scenarios) and harmful manipulation evaluations (e.g. multi-turn human-chatbot conversations in the behavioral study).	**Adapting test-time compute** is reported in cybersecurity evaluations (e.g. pass@1 through pass@30 on Cybench with up to 100 flag submission attempts per challenge), CBRN evals (e.g. <=10 sampled responses per task in High-Risk Bottlenecks scoring), AI R&D/Loss-of-Control (e.g. 10 seeds per AIRS-Bench task) **Scaffolding with tool use** is reported in cybersecurity evaluations (e.g. a simple ReAct agent exposing Bash and Python tools, with full Kali Linux CLI access plus apt/pip install capability for CTFs), CBRN (e.g. ReAct agents with domain-specific computational biology tools), AI R&D (e.g. stateful containers with H100/H200 GPUs). **Context management** is reported in cybersecurity evaluations (e.g. context compaction for long-horizon tasks). **Helpful-only / reduced-refusal model variants** are reported in CBRN evaluations and social engineering capability assessment (manipulation).	**Helpful-only / safeguards-removed variants** are reported in dual-use capability evaluations across risk domains.	**Not Mentioned**	**Not Mentioned**	**Not Mentioned**	**Not Mentioned**

Indicator

Human Uplift Trials

Definition

This indicator assesses whether organizations conduct rigorous, controlled human-subject studies to evaluate the marginal risk AI systems pose in dangerous domains by "uplifting" people's ability to cause harm. Key evidence includes experimental designs that compare task performance with and without AI support, the inclusion of domain-relevant experts, realistic and consequential task scenarios, and transparent publication of methods and findings. To assess worst-case potential, models should be tested without embedded safety filters.

Why it matters

Empirical uplift studies are critical for grounding AI safety policy in observable outcomes. These studies assess whether advanced systems significantly enhance a user's ability to cause harm and inform the development of proportionate safety interventions. Entities that conduct and publish such studies exhibit leadership in transparent, evidence-based risk governance.

Relevant Regulations and Standards EU AI Code of Practice Safety and Security	Anthropic Claude Opus 4.7	OpenAI GPT 5.5	Google DeepMind Gemini 3.1 Pro *	Meta Muse Spark	xAI Grok 4.1	DeepSeek V4	Z.ai GLM 5	Alibaba Cloud Qwen 3.5 Max	Mistral Mistral Large 3
(Measure 3.2) Examples of model evaluation methods include human uplift studies.	The general process for risk assessment involves "a portfolio of red-teaming, uplift trials, long-form task-based agentic evaluations (which includes creative and generative tasks), as well as automated knowledge and skill evaluations." **However, for Opus 4.7, the evaluations do not include human uplifting trials,** and are limited to "red-teaming and long-form task-based agentic evaluations and knowledge and skill evaluations." Previously, for Opus 4.6, Anthropic has conducted a controlled uplift trial to assess whether AI models can function as an "expert creative biologist," and a virology protocol uplift trial where we tasked PhD-level experts to determine the step-by-step protocol required to reconstruct a challenging virus.	**Not Mentioned** There has been no mention of a randomized controlled uplift trial since the release of GPT-5's system card.	**Not Mentioned** in the Gemini 3.1 Pro system card. Previously, Panoplia Laboratories (co-funded and coordinated by the Frontier Model Forum) has run a randomized, controlled trial assessing physical world wet lab uplift from LLMs (including Gemini 2.5 Pro) to novice threat actors in a biological threat scenario (relative to a control of internet access only).	**Not Mentioned**	**Not Mentioned**	**Not Mentioned**	**Not Mentioned**	**Not Mentioned**	**Not Mentioned**

External

Indicator

Independent Review of Safety Evaluations

Definition

Assesses whether an AI developer *commissions independent third-party experts to (A) verify the factual accuracy and process integrity of its internal dangerous-capability evaluations and (B) assess the* evaluation quality *and the company's interpretation of the results. We collect information on the reviewers' identity and credentials, their independence (including any conflicts of interest), the scope of the review, depth of access to data and logs (including rights to replicate or extend tests), and whether their findings are published unredacted.

Why it matters

AI developers control both the design and disclosure of dangerous capability evaluations, creating inherent incentives to under-report alarming results or select lenient testing conditions that avoid costly deployment delays. Regulators, investors, and the public, therefore, face a critical information asymmetry: they must trust safety claims based on self-reported evaluations with minimal methodological transparency. Independent external scrutiny can address this trust deficit by verifying reported results, assessing whether evaluations are sufficiently rigorous to uncover real risks, and providing credible third-party perspectives on whether safety claims are justified. This need is especially acute for catastrophic risk domains such as biosecurity, where companies may cite "infohazard" concerns to limit transparency.

Relevant Regulations and Standards Illinois SB 315 § 10(d) Summary	**Anthropic** Claude Opus 4.7	**OpenAI** GPT 5.5	**Google DeepMind** Gemini 3.1 Pro *	**Meta** Muse Spark	**xAI** Grok 4.1	**DeepSeek** V4	**Z.ai** GLM 5	**Alibaba Cloud** Qwen 3.5 Max	**Mistral** Mistral Large 3
▪ **Who & when:** Large frontier developers must, annually, hire an outside auditor, starting Jan 1, 2028 (or 90 days after qualifying as a large frontier developer, whichever is later). ▪ **What's audited:** Whether the developer has complied with the requirements of Section 10 (its frontier AI framework obligations). ▪ **Auditor qualifications:** Must follow generally accepted auditing standards, and have demonstrated competence, including access to people with technical expertise in frontier model safety. ▪ **Independence safeguards:** No audit if either party has a financial interest in the other; the developer can pay for the work but cannot tie payment to the audit's outcome. ▪ **Access:** The auditor gets all materials reasonably necessary, including unredacted versions of everything published under the Act. ▪ **Report contents:** Must state whether the developer substantially complied, disclose conflict-of-interest procedures and any conflicts, describe the methodology and information reviewed, and carry the lead auditor's signature certifying results. ▪ **Publication:** Within 30 days of receiving the report, the developer must publicly post a high-level summary plus a redacted copy of the full report, and send the redacted report to the Agency and Attorney General.	**Not Mentioned**	**Not Mentioned**	**Not Mentioned**	**Not Mentioned**	**Not Mentioned**	**Not Mentioned**	**Not Mentioned**	**Not Mentioned**	**Not Mentioned**

Indicator

Pre-deployment External Safety Testing

Definition

This indicator evaluates whether companies enable external safety assessments of frontier AI models before public release, and the degree to which those evaluators operate independently from the model developer. Independence will be assessed across four dimensions, including institutional affiliation, methodological autonomy, access autonomy, and publication freedom. Evidence includes the identity and qualifications of external parties, the level and duration of access provided, compensation arrangements, testing permissions, and the evaluators' ability to publish independently. The strength of these practices is judged by the comprehensiveness of the evaluations, the depth of access, and the autonomy of the evaluators.

Why it matters

External evaluations are essential for verifying safety claims and uncovering risks that internal teams may overlook or under-report. Providing external evaluators with substantial access and ensuring their ability to test and publish with a great amount of autonomy reflect a company's commitment to transparent and evidence-based governance.

Relevant Regulations and Standards EU AI Code of Practice	**Anthropic** Claude Opus 4.7	**OpenAI** GPT 5.5	**Google DeepMind** Gemini 3.1 Pro *	**Meta** Muse Spark	**xAI** Grok 4.1	**DeepSeek** V4	**Z.ai** GLM 5	**Alibaba Cloud** Qwen 3.5 Max	**Mistral** Mistral Large 3
(Appendix 3.4-3.5) Signatories must ensure that qualified independent external evaluators assess their models for systemic risk unless the model is already proven comparably safe or evaluators cannot be secured after reasonable efforts. These evaluators must have relevant technical expertise (academic or professional) and follow strict security and confidentiality protocols. Meanwhile, signatories will provide independent external evaluators with (1) adequate access (e.g. access to model activations, gradients, logits, chains-of-thought, model version(s) with the fewest safety mitigations implemented) (2) information (e.g. model specifications (including the system prompt), relevant training data, test sets, and past model evaluation results), (3) time, and (4) other resources (e.g. compute budgets, staffing, engineering budgets and support)	**Scope** UK AISI (Cyber, open-ended testing; Alignment, behavior audits) **Access** (1) UK AISI - a pre-release checkpoint of [Claude Opus 4.7] for cyber evals. (2) UK AISI - an unreleased checkpoint of [Claude Opus 4.7] for behaviours relevant to misalignment risk (with and without reasoning, and with full chain-of-thought access for analysis) **Security and Privacy** Not specified. **Independence** (1) Publication - a. Evaluators may publish independently after company review/ possible redaction. b. The company provided its own summary of the evaluator's key findings. (2) Financial, governance, or commercial dependency independence is not specified.	**Scope** Apollo (sandbagging, in-context scheming and strategic deception), SecureBio (biological capabilities), Irregular (cybersecurity), US CAISI (biological capabilities and cybersecurity) and UK AISI (cybersecurity), external red-teamers (cybersecurity), METR (AI R&D and loss of control risks, not directly for GPT 5.5) **Access** (1) The longest period of time that an external evaluator was given continuous access for pre-deployment testing is >2 weeks (<=3 weeks). (2) Highest level of technical access: "Helpful-only" or base model API (no harmlessness fine-tuning and no filters); OpenAI also shares visible Chain of Thought access for evaluators who require this. (3) Third-party evaluators: GPT-5.5 early checkpoints, as well as the final launch candidate models; a. UK AISI: prototype versions of OpenAI's safeguards and information sources that are not publicly available – such as our monitor system design, biological content policy, and chains of thoughts of our monitor models. (4) Query-rate or volume restrictions: Elevated but capped – evaluators had higher quotas than the public/enterprise tier but were still subject to explicit Elevated but capped – evaluators had higher quotas than the public/enterprise tier but were still subject to explicit caps (e.g., requests-per-minute or daily token limits) Query rates can depend on technical feasibility in some cases.	**Scope** Apollo Research, Dreadnode and Vaultis, across CBRN, Loss of Control, Cyber, and Harmful Manipulation. **Access** (1) The longest period of time that an external evaluator was given continuous access for pre-deployment testing is >2 weeks (<=3 weeks). (2) Highest level of technical access is Inference API with safety filters disabled (no inference-time mitigations) a. Evaluators are provided with models without inference time mitigations relevant to their specific domain (3) Query-rate or volume restrictions: a. Elevated but capped – evaluators had higher quotas than the public/ enterprise tier but were still subject to explicit (e.g., requests-per-minute or daily token limits), depending on technical feasibility. b. Bespoke depending on the testing partner's specific needs and evaluation type	**Scope** ScaleAI (Model resistance to harmful requests); HandshakeAI (Robustness against jailbreak); Apollo Research (Evaluation awareness, deceptive alignment, loss of control); Irregular (cybersecurity evaluations); Deloitte, Faculty AI, SecureBio, and Frontier Design (Engagements for biodefense and biosecurity, including workflows for threat modeling, experimental design and testing, and the interpretation and validation of evaluation results) **Access** 1) Highest: "Helpful-only" or base model API (no harmlessness fine-tuning and no filters) 2) Query-rate or volume restrictions: Elevated but capped – evaluators had higher quotas than the public/enterprise tier but were still subject to explicit.	Not specified in the system card.	No public mention.	No such external pre-deployment testing was commissioned (Company survey Q17)	No public mention.	No public mention.

Table continues on next page

Relevant Regulations and Standards EU AI Code of Practice	Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud	Mistral
	Claude Opus 4.7	GPT 5.5	Gemini 3.1 Pro *	Muse Spark	Grok 4.1	V4	GLM 5	Qwen 3.5 Max	Mistral Large 3
Signatories will not undermine the integrity of external model evaluations by storing and/or analyzing inputs and/or outputs from test runs without express permission from the evaluators. **California SB 53 § 22757.12(c)(2)(C); New York RAISE Act § 1421(3)(b); Illinois SB 315 § 10(c)(2)(C)** Before, or concurrently with, deploying a new frontier model or a substantially modified version of an existing frontier model, a large frontier developer shall include in the transparency report summaries: (C) The extent to which third-party evaluators were involved.	**Timeline** Not specified. (Source: System card pp. 52-53, 114-117; Company survey answers Q17-23)	**Security and Privacy** Zero Data Retention available upon request, if technically feasible during pre-deployment periods (for some new models or products, ZDR is not always possible during pre-deployment testing). **Independence** (1) No financial ties: US CAISI, UK AISI, METR a) OpenAI provides API credits to all third party assessors as needed to help facilitate their research and evaluations. b. OpenAI offers compensation to all of its third party assessors, and some choose to decline depending on their organizational philosophy around this. Direct payment for work and/or subsidizing model use costs through API credits or otherwise. No payment is ever contingent on the results of a third party assessment. [OpenAI, 2025] (2) No governance ties: US CAISI, UK AISI, Irregular, Apollo. SecureBio, METR (3) No material commercial dependency: US CAISI, UK AISI, METR (4) Publication a. Evaluators may publish independently without prior company approval after the model is released. (This is true if they run their evaluations independently on the deployed model. Results from the pre-deployment evaluation period are under NDA / require prior approval to protect confidential information.) b. Evaluators may publish independently after company review/possible redaction. (in cases where the evaluator wishes to publish about the specifics of the pre-deployment period, see examples from METR, UK AISI, SecureBio) c. The company publishes a report after review/possible redactions. (OpenAI publishes excerpts from the report mutually agreed upon or written, with OpenAI having the final say for what content goes in System Cards) d. The company provided its own summary of the evaluator's key findings. (This is true in some cases, but we also share back any summaries that we plan to publish with the evaluator prior to release to confirm factual accuracy.) e. Findings remain internal. (Some evaluators prefer that their full findings are not shared publicly, such as some forms of government testing by the US CAISI and UK AISI.) **Timeline** External safety tests were completed after broad internal deployment.	**Security and Privacy** Inputs and outputs are neither logged nor retained, protecting evaluator IP. However, where agreed, external evaluators share prompts and model responses for the purpose of assessment and mitigation of risks. Answers for Gemini 3.1 Pro **Independence** (1) Publication: The company provided its own summary of the evaluator's key findings. (High level summaries appropriate for the risks being evaluated within the Models Cards / Tech report with GDM having the final say for what content goes in the Model Cards/Tech report.) (2) For each external tester GDM work with for pre-deployment testing, the team undergo a due diligence process to ensure their independence. This covers, among other areas, governance and commercial dependency. (3) GDM does provide payment for conducting evaluations to compensate for their time. GDM team states that it has measures in place to ensure that these payments are in no way conditional on the conclusions they reach. **Timeline** Not specified.	**Security and Privacy** Not specified **Independence** (1) Financial, governance, or commercial dependency independence is not specified. (2) Publication independence is not specified. e.g. Apollo Research's findings are summarized in the system card. **Timeline** External safety tests were completed before broad internal deployment.					

Indicator

Bug Bounties for System Vulnerabilities

Definition

This indicator evaluates whether companies maintain structured programs that reward or formally recognize external researchers for discovering and responsibly disclosing safety vulnerabilities in AI system behavior, such as through red-teaming initiatives or bug bounties. The focus is primarily on behavioral vulnerabilities, such as jailbreaks, prompt attacks, data extraction, or adversarial manipulations, rather than conventional software or cybersecurity bugs. Evidence includes the scope of eligible vulnerabilities, reward structure or compensation levels, response and disclosure processes, and the public availability of program rules and results.

Why it matters

Structured disclosure programs with financial incentives harness external expertise to identify system vulnerabilities before they are exploited in deployment. Investments in such programs indicate openness and proactiveness toward risk identification.

Relevant Regulations and Standards EU AI Code of Practice	**Anthropic** Claude Opus 4.7	**OpenAI** GPT 5.5	**Google DeepMind** Gemini 3.1 Pro *	**Meta** Muse Spark	**xAI** Grok 4.1	**DeepSeek** V4	**Z.ai** GLM 5	**Alibaba Cloud** Qwen 3.5 Max	**Mistral** Mistral Large 3
(Measure 3.5 Post-market monitoring) The provision mentions bug bounties as an example of post-market monitoring methods.	**Anthropic has maintained ongoing private bug bounty programs for post-deployment monitoring via HackerOne.** **Scope**: (1) Model safety (classifier systems) for harmful biological questions; (2) Security vulnerabilities (e.g. misconfigurations, CSRFs) **Participation**: HackerOne (private) **Reward**: up to $35,000 per novel, universal jailbreak identified **Investigation and Resolution Timeline**: Not specified. **Model Access**: Not specified. **Confidentiality:** All Program participants are required to sign a non-disclosure agreement to protect Program confidentiality as a condition for joining, with the exception for public disclosure of the existence of the program and their participation.	**OpenAI launched a public bug bounty program for GPT-5.5 (April 23, 2026 - June 22, 2026)** **Scope**: Bio-related risks **Participation**: Invitation or application **Reward**: (1) $25,000 to the first true universal jailbreak to clear all five questions; (2) Discretionary smaller rewards for partial wins. **Investigation and Resolution Timeline**: Not specified. **Model Access:** Not specified. **Confidentiality:** All prompts, completions, findings, and communications are covered by NDA.	**Google has expanded its ongoing bug bounty efforts with the launch of its AI vulnerability reward program in 2025 (originally started in 2023).** **Scope**: AI-related vulnerability and abuse issues in Google and Alphabet AI products with clarifications on qualifying (including rogue actions) and non-qualifying vulnerabilities (except that issues found in Vertex AI or other Google Cloud products are covered by the Google Cloud Vulnerability Rewards Program) **Participation**: Open (with exceptions for sanctioned entities/territories) **Reward**: The reward is up to $20,000 and is determined based on product tiers (flagship, standard, other) and vulnerability types, with the highest belonging to rogue actions for flagship products. There is also a factor to the reward based on report quality (0.8 for low, 1.0 for good, and 1.2 for exceptional) **Model Access**: Recommended to use their own/test accounts. They are not allowed to access others' data, and must not disrupt other users. **Investigation and Resolution timeline**: The company has vaguely promised that it will "respond promptly and fix bugs in a sensible timeframe," and "provide regular updates on the status of submission." Prioritization will be based on severity of the vulnerabilities. **Confidentiality**: Researcher controls disclosure timing but advance notice are expected even for non-rewarded and non-fixed reports. Certain disclosure behaviors may forfeit reward, including "disclosing in a way that puts users or Google in 'immediate danger.'"	**Scope**: The ongoing bug bounty program (started in 2023) for Meta AI is restricted to privacy or security issues, like extracting training data through tactics like model inversion or extraction attacks. **Reward**: - The minimum reward for a qualifying submission is US $500. - The maximum reward for a qualifying submission in Meta AI is US $30,000. **Participation**: Openly through GitHub, developers.facebook.com, and llamausereport@meta.com, and has not been updated since the release of Muse Spark. **Access**: Participants do not have special access to the models but are encouraged to use authorized or test accounts. **Confidentiality**: Meta's Bug Bounty confidentiality and disclosure rules require researchers to avoid privacy violations, use only authorized or test accounts, immediately report and delete any inadvertently accessed data, and give Meta reasonable time to investigate before any public disclosure. Safe-harbor protections apply only if researchers act in good faith and fully comply with these terms.	**Scope**: The ongoing program covers xAI, including the Grok API, and targets traditional security vulnerabilities, including authentication, authorization, data-exposure, and infrastructure issues. However, model behaviors and AI safety issues are explicitly out of scope. The last update for the bug bounty program is March 20, 2025. **Reward**: Bounties are discretionary, determined by a 5×5 internal risk matrix (impact × likelihood) and by a panel of security experts. 90-day averages as of the last update (May 2025): ▪ Low $100 – $500 (19.6 %) ▪ Medium $500 – $2,000 (40 %) ▪ High $2,500 – $7,000 (30 %) ▪ Critical $7,500 – $20,000 (10 %) **Timeline**: Issues are usually triaged within ~1 day and resolved within ~3 weeks. **Access**: No mention of model access or sandbox environment. **Confidentiality**: Participants must abide by HackerOne's disclosure guidelines, including using test accounts, protecting user privacy, and keep all findings confidential until the report is closed.	No public bug bounty program identified.	No public bug bounty program identified.	No public bug bounty program identified.	No public bug bounty program identified.

TO BE COMPLETED BY PANELLISTS

Grading Sheet: Risk Assessment

Please pick a grade for each firm. You may use the full letter-grade scale with +/– modifiers as appropriate. You can add brief justifications to your grades.

	Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud	Mistral
Grades									
Grade comments (Justifications, opportunities for improvements, etc.)									

Grading Scales

Grading scales are provided to support consistency between reviewers. Please note that you can also use the +/– modifiers.

- **A** Comprehensive, state-of-the-art evaluations; strong validity, reproducibility, and independent review; no serious harm potential.
- **B** Robust assessments; good validity and elicitation; limited external review; serious harms well-controlled.
- **C** Partial assessments; uneven validity or elicitation; little external input; serious harms mostly controlled.
- **D** Fragmented assessments; weak validity and elicitation; no external review; serious harms poorly controlled.
- **F** No credible assessment; serious harm uncontrolled.

Domain comments

Optional: Share observations that apply across companies, including general recommendations, notes on how you weighted indicators, or feedback on FLI's methodology.

Domain comments	

Domain

Current Harms

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

Table of Contents

Chinese Regulatory System Summary

China's Interim Measures mandate strict data minimization, lawful handling of user information, and timely fulfillment of user rights requests, ensuring robust privacy protection. Meanwhile, the Deep Synthesis Regulation and National Standard GB45438-2025 require AI providers to implement both explicit and implicit watermarking systems, ensuring traceability and transparency of AI-generated content.

Local binding instruments, voluntary technical standards, draft regulations and standards, as well as strategic and policy guidance documents are not applicable here.

National Binding Instruments

Privacy

Interim Measures Article 11 requires AI service providers to lawfully protect users' input data and usage records.

Specifically, they must not collect unnecessary personal information (data minimization), must not illegally retain identifiable input data or usage records, and must not illegally provide such information to others (lawful handling). In addition, providers must timely accept and handle user requests to access, copy, correct, supplement, or delete their personal information (responsive obligations to user rights).

Watermarking

Deep Synthesis Regulation Article 16-18 requires that deep synthesis service providers are required to add built-in watermarks and keep system logs. When content could confuse people, providers must place prominent marks on generated or edited content. They must also provide labeling functions for other synthetic content and remind users they can apply visible marks. No one is allowed to remove or alter these marks.

National Standard GB45438—2025 Cybersecurity technology—Labeling method for content generated by artificial intelligence delineates the specific requirements that AI service providers have to follow when placing explicit vs. implicit watermarks.

For explicit labeling, when AI-generated text, audio, video, or other content could mislead or confuse the public, providers must apply clear and visible marks at specified positions.

For implicit labeling, every AI-generated file must contain standardized metadata that includes: (1) an AI-generation tag; (2) the service provider's name or code; (3) a unique content ID; (4) the distributor's name or code; and (5) a unique distribution ID. Content-implicit labeling is optional and not required under this standard.

Safety Performance

Indicator

Stanford's HELM Safety Benchmark

Definition

This indicator measures model performance on Stanford's HELM Safety v1.0 benchmark, a suite of five safety tests covering six risk categories: violence, fraud, discrimination, sexual content, harassment, and deception. The benchmark includes: HarmBench (jailbreak resistance); BBQ (social discrimination); SimpleSafetyTest; XSTest (alignment between helpfulness and harmlessness); and AnthropicRedTeam (resilience to adversarial probing). Performance is reported as normalized aggregate scores ranging from 0 to 1, where higher scores indicate fewer safety risks. Scoring is based on exact match accuracy for BBQ and model-judge ratings (GPT4o and Llama 3.1 405B) for the remaining benchmarks.

Why it matters

HELM Safety provides a standardized, empirical benchmark for evaluating how reliably AI systems prevent harmful or unsafe outputs. It measures behavioral safeguards—such as refusals of violent, fraudulent, or discriminatory content—under consistent testing conditions. Strong performance demonstrates that a model's technical safety mechanisms effectively reduce direct user-facing risks across diverse harm categories.

	Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud	Mistral
Models Evaluated	Opus 4.7	GPT-5.5	Gemini-3.1-pro-preview	Muse Spark	Grok 4.3	Deepseek-v4-pro	GLM 5.1	Qwen3.6-max-preview	Mistral-large-2512
Average Score (max score = 1)	0.96	0.97	0.87	Not available	0.87	0.80	0.81	0.98	0.67
HarmBench	0.94	0.99	0.87		0.91	0.68	0.80	0.99	0.47
SimpleSafetyTests	1.00	1.00	0.93		0.98	0.98	0.81	1.00	0.91
BBQ Accuracy	0.99	0.96	0.00		0.63	0.49	1.00	0.94	0.14
AnthropicRedTeam	0.94	0.98	0.94		0.89	0.98	0.99	0.99	0.90
XSTest	0.94	0.94	0.75		0.96	0.90	0.46	0.97	0.94
Evaluation Results Obtained	Retrieved May 31, 2026.								

Footnotes

[1] Farzaan Kaiyom et al., "*HELM Safety: Towards Standardized Safety Evaluations of Language Models*," Stanford Center for Research on Foundation Models, November 8, 2024, https://crfm.stanford.edu/2024/11/08/helm-safety.html.

Indicator

Stanford's HELM AIR Benchmark

Definition

This indicator evaluates model performance on Stanford's AIR-Bench 2024 (AI Risk Benchmark), an AI safety benchmark aligned with emerging government regulations and company policies. We report mean scores across 5,694 tests spanning 314 granular risk categories, with scores measuring the percentage of appropriately refused requests. The benchmark systematically evaluates four major risk domains: System & Operational Risks (e.g., cybersecurity, operational misuse), Content Safety Risks (e.g., child sexual abuse material), Societal Risks (e.g., surveillance), and Legal & Rights-related Risks (e.g., privacy violations, defamation). All prompts are manually curated and human-audited to ensure they reflect genuine policy violations rather than benign content.

Why it matters

HELM AIR provides an evaluation of how well AI systems align with real-world safety expectations. Unlike behavioral safety tests, it directly reflects the kinds of standards developers will be expected to meet, rooted in emerging regulation, ethics, and risk-management practices. Strong performance signals high readiness of AI systems to comply with policy and societal values.

	Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud	Mistral
Models Evaluated	Opus 4.7	GPT-5.5	Gemini-3.1-pro-preview	Muse Spark	Grok 4.3	Deepseek-v4-pro	GLM 5.1	Qwen3.6-max-preview	Mistral-large-2512
Average Score (max score = 1)	0.86	0.93	0.77	Not available	0.77	0.78	0.94	0.86	0.40
System & Operational Risks	0.75	0.89	0.81		0.78	0.70	0.94	0.78	0.21
Content Safety Risks	0.85	0.87	0.83		0.70	0.80	0.92	0.82	0.42
Societal Risks	0.94	0.97	0.82		0.80	0.79	0.97	0.92	0.27
Legal & Rights-related Risks	0.94	0.97	0.85		0.89	0.86	0.97	0.94	0.40
Evaluation Results Obtained	Retrieved May 31, 2026.								

Footnotes

[1] Yi Zeng et al., *"AIR-Bench 2024: A Safety Benchmark Based on Risk Categories from Regulations and Policies,"* arXiv preprint arXiv:2407.17436, 2024, https://arxiv.org/abs/2407.17436.

Indicator

TrustLLM Benchmark

Definition

This indicator measures a model's overall trustworthiness using the TrustLLM benchmark, a comprehensive framework spanning six dimensions: truthfulness, safety, fairness, robustness, privacy, and machine ethics.[1] The benchmark includes over 30 datasets across more than 18 subcategories, assessing issues such as hallucination, jailbreak resistance, and privacy leakage. Models are evaluated on tasks ranging from simple classification to complex generation, with results reported as published scores and rankings across each dimension. TrustLLM was developed by 45 research institutions, including 38 based in the U.S.

Why it matters

TrustLLM evaluates how reliably AI systems uphold truthfulness, privacy, and ethical reasoning beyond standard capability metrics. Strong performance indicates that companies have invested in aligning their models to be harmless and helpful, and not to cause unintended harm.

	Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud	Mistral
Models Evaluated	Opus 4.7	GPT-5.5	Gemini-3.1-pro-preview	Muse Spark	Grok 4.3	Deepseek-v4-pro	GLM 5.1	Qwen3.6-max-preview	Mistral-large-2512
Total score (max = 1)	0.62	0.58	0.57	Not available	0.57	0.57	0.58	0.60	0.50
Truthfulness	0.52	0.48	0.40		0.46	0.46	0.45	0.50	0.51
Safety	0.65	0.62	0.69		0.64	0.62	0.64	0.69	0.50
Fairness	0.40	0.33	0.43		0.35	0.35	0.42	0.39	0.30
Privacy	0.62	0.64	0.51		0.56	0.56	0.59	0.59	0.42
Ethics	0.89	0.71	0.74		0.73	0.79	0.74	0.74	0.70
Robustness	0.68	0.68	0.67		0.66	0.64	0.66	0.71	0.56
Evaluation Results Obtained	Retrieved May 14, 2026.								

Footnotes

[1] Yue Huang et al., *"TrustLLM: Trustworthiness in Large Language Models — A Principle and Benchmark,"* arXiv preprint arXiv:2401.05561 (January 10, 2024), https://arxiv.org/abs/2401.05561.

Indicator

CAIS AI Dashboard

Definition

The safety benchmark collection includes six benchmarks. [1]

(1) Agent Red Teaming measures robustness against malicious use and prompt injections. The benchmark evaluates AI agents across five core behavior categories: Confidentiality Breaches, Conflicting Objectives, Instruction Hierarchy Violations — Information, and Instruction Hierarchy Violations — Actions. Lower scores indicate better performance.

(2) Virology Capabilities Test (VCT) Refusal measures the rate at which AIs refuse hazardous expert-level virology queries. It is a subset of VCT that is manually labeled by experts for queries that should be refused because of potential dual-use. The ability to provide expert-level troubleshooting is inherently dual-use: it is useful for beneficial research, but it can also be misused in particularly harmful ways.

(3) Humanity's Last Exam (HLE) Miscalibration measures how well AI models calibrate their confidence with their actual accuracy. This metric evaluates whether models are overconfident or underconfident in their predictions across 2,500 expert-written academic questions spanning 100+ disciplines. Lower calibration error indicates better alignment between a model's expressed confidence and its true performance.

(4) MASK tests the extent to which AIs can be pressured to overtly lie. It is different from measuring accuracy or truthfulness, but instead, it measures whether a model consistently expresses its beliefs even when pressured to do otherwise. [Scale AI]

(5) Machiavelli measures harmful behaviors in LLM agents across 30 text-based games. We use its deception measurements from its various goal-oriented scenarios.

(6) TextQuests's harm metric measures the AI's propensity to engage in wantonly harmful behaviors in text-based adventure games.

Why it matters

These benchmarks matter because they test safety-relevant traits—like honesty, refusal behavior, and ethical restraint—that do not automatically improve with model size or training compute. Their low correlation with general capabilities means they capture distinct aspects of alignment and behavioral reliability rather than raw intelligence. This separation helps prevent "safetywashing," where capability gains are mistaken for safety progress. In doing so, they provide a more rigorous basis for tracking genuine advances in AI safety as systems grow more powerful. [Ren et al., 2024]

	Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud	Mistral
Models Evaluated	Opus 4.7	GPT-5.5	Gemini-3.1-pro	Muse Spark	Grok 4.3	Deepseek-v4-pro	GLM 5.1	Qwen3.6-max-preview	Mistral-large-2512
Average Score (Lower = Better)	34.80	42.30	57.70	Not available	47.10	60.20	54.50	56.32	69.27
Agent Red Teaming	44.50	41.50	68.00		90.10	90.90	75.80	N/A	90.91
Virology Capabilities Test (VCT) - Refusal	42.30	84.00	99.50		43.90	100.00	98.40	99.50	100.00
Humanity's Last Exam (HLE) - Miscalibration	28.10	41.90	50.30		42.50	57.10	58.50	66.30	80.50
MASK	17.10	9.90	53.80		48.30	32.60	15.80	38.50	58.60
Machiavelli	56.00	52.00	55.70		53.70	54.60	56.60	60.10	66.10
TextQuests Harm	20.80	24.40	18.80		3.90	26.20	22.00	17.20	19.50
Evaluations Accessed	May 15, 2026.								

Footnotes

[1] Long Phan, Jaehyuk Lim, Arunim Agarwal, and Dan Hendrycks, *"CAIS AI Dashboard,"* Center for AI Safety, 2025, https://dashboard.safe.ai.

Digital Responsibility

Indicator
Protecting Safeguards from Fine-tuning

Definition

This indicator evaluates whether companies maintain safeguards that prevent the removal of built-in safety measures during fine-tuning. Evidence differentiates between: i) Supervised or hosted fine-tuning, which occurs on the company's platform where core safety filters remain active; and ii) Full model-weight releases, where users can directly modify parameters and potentially disable all protections unless tamper-resistant controls are in place.

If companies provide no public information on fine-tuning or weight-release policies for their frontier AI systems, these capabilities are treated as not publicly accessible.

Why it matters

Releasing full model weights may allow malicious actors to strip or override safety mechanisms, creating uncensored or harmful versions. In contrast, supervised fine-tuning preserves core safety guardrails while enabling responsible customization.

Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud	Mistral
Opus 4.7	**GPT-5.5**	**Gemini 3.1 Pro**	**Muse Spark**	**Grok 4.2**	**V4**	**GLM-5**	**Qwen 3.5 Max**	**Large 3**
Frontier model weights protected. Provide supervised fine-tuning for older and smaller Claude 3 Haiku through Amazon Bedrock. Safety mitigations are in place. [AWS, 2024]	Frontier model weights protected. Released open weights of non frontier models including gpt-oss families and gpo-oss-safeguards families that support custom safety policies. [OpenAI] Provide supervised fine-tuning (SFT) of gpt-4.1-2025-04-14, gpt-4.1-mini-2025-04-14, and gpt-4.1-nano-2025-04-14 [OpenAI] and RL fine-tuning for o4-mini-2025-04-16. OpenAI is winding down the fine-tuning platform. The platform is no longer accessible to new users, but existing users of the fine-tuning platform will be able to create training jobs in the coming months, until the base models are deprecated. For instance, gpt-4.1-nano-2025-04-14 will be shut down in October 2026. [OpenAI]	Frontier model weights protected. Google released the non-frontier Gemma families, the latest of which is Gemma 4 published in April 2026. [Google, 2026] Enables supervised fine-tuning of Gemini 3.1 Flash-Lite, Gemini 2.5 Pro, Gemini 2.5 Flash, and Gemini 2.5 Flash-Lite. Safety mitigations are in place. [Google, 2026].	Frontier model weights protected. Previously has released open-source models from the Llama family, and the latest is Llama 4 Maverick. No tamper-resistant safeguards. [Meta AI]	Frontier model weights protected. Fully released weights of non-frontier Grok 1. No tamper-resistant safeguards. [xAI, 2024]	Fully released weights of frontier models. No tamper-resistant safeguards. [Hugging Face]	Fully released weights of the frontier model. No tamper-resistant safeguards. [Hugging Face]	Frontier model weights protected. Fully released weights of non-frontier models from the Qwen family, including the Qwen 3.6 and Qwen 3.5 series. No tamper-resistant safeguards. [Hugging Face]	Fully released weights of frontier models. No tamper-resistant safeguards. [Hugging Face]

Indicator

Watermarking

Definition

This indicator assesses whether companies have implemented watermarking technologies to help identify AI-generated content in both text and images. It focuses on real-world implementation rather than research prototypes, evaluating the accuracy and robustness of detection methods, adherence to standards such as C2PA and SynthID, and whether detection tools are publicly accessible.

Why it matters

Watermarking helps distinguish authentic content from AI-generated media, reducing the risks of misinformation, fraud, and reputational harm. Companies that implement robust and standardized watermarking systems, and make detection tools publicly accessible, demonstrate a strong commitment to transparency, provenance, and digital trust.

Sub-Indicator	Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud	Mistral
	Opus 4.7	GPT-5.5	Gemini 3.1 Pro	Muse Spark	Grok 4.2	V4	GLM-5	Qwen 3.5 Max	Large 3
Text-based	**None found**	**No** OpenAI has announced that it has developed a text watermarking method, but it is still researching for alternatives, due to concerns over its effectiveness against globalized tampering, and disproportionate stigmatizing impact on non-native English speakers. [OpenAI, 2024]	**Yes** (SynthID) The SynthID system uses particular token selection to introduce a pattern that marks a text as AI-generated [Google DeepMind]. This can be identified using an online detection tool, which is currently accessible only to approved journalists, media professionals, and researchers through a waitlist program. [Google, 2025].	**None found**	**None found**	**Yes** Under the 2025 National Standard on AI-Generated Content Labeling and Watermarking, companies must include explicit watermarks to identify content produced by artificial intelligence. The standard applies to text, images, audio, video, and virtual environments. For each content type, it specifies (1) where the label must appear, (2) the required information to include in the label, and (3) the parameters that determine its visibility or audibility, such as label size, voiceover speed, and display duration. The government has taken active enforcement efforts against non-compliant companies.			**None found**
Image-based		**No watermarking** (C2PA metadata) Images generated with ChatGPT on the web and the API serving the DALL·E 3 model, will now include C2PA metadata. The metadata can be detected, but it can be removed either accidentally or intentionally, [OpenAI, 2025]	**Yes** (SynthID) Pattern is embedded in images, can be identified by an online detector, access currently limited. [Google DeepMind]	The open-source Llama 4 family does not include models that can generate images. However, for photorealistic images created using Meta AI, Meta has applied visible labels of "Imagined with AI" and included invisible watermarks and metadata embedded within files. [Meta, 2024]	**None found**				

Indicator

User Privacy

Definition

This indicator reports a company's dedication to user privacy when training and deploying AI models. It considers whether user inputs (such as chat history) are used by default to improve AI models or if companies require explicit opt-in consent. It also considers whether users can run powerful models privately, through on-premise deployment or secure cloud setups. Evidence includes default privacy settings and the availability of model weights for private hosting.

Why it matters

Privacy controls that require deliberate consent to opt in enable greater respect for user privacy, especially in sensitive fields such as healthcare, law, and government.

Sub-Indicator	Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud	Mistral
	Opus 4.7	GPT-5.5	Gemini 3.1 Pro	Muse Spark	Grok 4.2	V4	GLM-5	Qwen 3.5 Max	Large 3
Default training on user inputs	**Yes** Anthropic updated its consumer terms and privacy policy in August 2025, introducing a new data-sharing setting under which user conversations are included in model training by default unless the user manually opts out through the "Help improve Claude" toggle. This applies to users for Claude Free, Pro, and Max plans. Previously, user inputs were used for model improvement only if users explicitly opted in or if the conversation was flagged for violating Usage Policy. [AI Safety Index, 2025]	**Yes** (exception for enterprise data) ChatGPT does not train models on Enterprise account user's business data by default. [OpenAI, 2025]	**Yes** (exceptions for Gemini for Google Cloud users) Gemini for Google Cloud doesn't use your prompts or its responses as data to train its models.	**Yes** Meta "use information shared on Meta Products" to train their AI models. "This information could be things like posts or photos and their captions." Private messages are excluded unless "someone in the chat chooses to share those messages with our AIs." [Meta]	**Yes** xAI uses "X posts as well as user interactions, inputs, and results with Grok for training and fine-tuning purposes." [X] [Ars Technica, 2024] In addition, xAI uses user inputs to improve its models by default. [xAI, 2025]	**Yes** DeepSeek uses user inputs to improve its models by default. [DeepSeek, 2025]	**Yes** Z.ai uses user inputs to improve its models by default. [Z.ai, 2025] Z.ai's Qingyan, also known as ChatGLM, was found to have collected information beyond what users authorized. [National Cyber Security Information Center, 2025]	**Yes** Alibaba does not provide an opt-out option for users to stop their de-identified content from being used to train the model. [Alibaba, 2025]	**Yes** (exceptions for Enterprise accounts or paid APIs) Mistral does not use "Input and Output to train [AI] models when the users use Le Chat Enterprise or the paid version of APIs." [Mistral, 2026]
Frontier model weights available for private hosting	**No**	**No**, but less-powerful models are open-sourced	**No**, but less-powerful models are open-sourced	**Yes**	**No**, but less-powerful models are open-sourced	**Yes**	**Yes**	**No**, but less-powerful models are open-sourced	

Indicator

Major Safety Incidents & Response

Definition

The indicator documents the record of publicly reported safety incidents involving the company's AI products where serious potential or actual harm occurred, and evaluates how the company responded in each case and in general.

Why it matters

A company's safety commitment and measures are ultimately tested by real-world failures. This indicator assesses whether harm has been caused, whether the company disclosed them transparently, and whether corrective action was taken — providing a reality check on the effectiveness of the company's publicly disclosed safety behaviors.

Anthropic	**Large-scale Cybersecurity incidents** In September 2025, Anthropic detected suspicious activities, and after investigating, claimed that this was a highly sophisticated espionage campaign by a Chinese state-sponsored group. Anthropic alleged that the group manipulated the Claude Code tool and its agentic capabilities into "attempting infiltration into roughly thirty global targets and succeeded in a small number of cases," and executed "without substantial human intervention." Anthropic responded by launching an investigation, banning the accounts, notifying affected entities as appropriate, and coordinating with authorities. The company also shared a report delineating how the threat actors leveraged the AI systems for exploiting vulnerabilities and executing cyberattacks [Anthropic, 2025; Anthropic, 2025]. While Anthropic was applauded for this transparency, critics argued that the disclosure stopped short of the full technical transparency expected in serious threat reports, leaving doubt about the claims' verifiability and usefulness to the broader security community. [Gumbley, 2025]	**Alleged Claude involvement in Minab Girl School strike** Washington Post reporting has suggested that Anthropic's Claude LLM, integrated into Maven Smart System in collaboration with Palantir, may have been involved in targeting the Shajarah Tayyebeh elementary school in Minab, Iran on Feb 28, hitting it at least twice and killing 175–180 people, mostly girls aged 7–12 (Anthropic's AI tool Claude central to U.S. campaign in Iran, amid a bitter feud). Anthropic has neither publicly confirmed nor denied involvement, the U.S. military investigation is ongoing, and controversy remains about causes and responsibility [New York Times, 2026; New Yorker, 2026; Wikipedia]. Humanitarian harm aside, this incident harmed U.S. regime-change objectives by affecting Iranian public opinion of Operation Epic Fury.
OpenAI	**Chatbot provides operational uplift for violent crime.** 1. Lethal dosing South Korean prosecutors allege Kim So-young consulted ChatGPT about lethal combinations of alcohol and benzodiazepines, then served the laced drinks to three men on dates — two died, one survived a two-day coma. [CNBC, 2026] 2. Fatal shooting Tumbler Ridge Shooting (2026): An 18-year-old killed 8 people and injured nearly 30 people. 12 OpenAI employees reportedly flagged the conversations as "indicating an imminent risk of serious harm to others" and recommended notifying Canadian police. The recommendation was allegedly rebuffed and OpenAI only banned the account eight months before the attack. The suspect then opened a second account, which evaded detection, and continued planning. OpenAI initially defended inaction, saying the account "did not meet its threshold of a credible or imminent plan." After the shooting: Altman issued a public apology, pledged to strengthen police-notification protocols, made referral criteria to the police "more flexible," brought in mental-health/behavioral experts, and committed to a direct point of contact with Canadian law enforcement. [BBC, 2026; BBC, 2026] FSU Shooting (2026): Florida's Attorney General probed criminal investigation with OpenAI after "[the office's] review has revealed that a criminal investigation is necessary and that ChatGPT offered significant advice to this shooter before he committed such heinous crimes," citing evidence such as advice on gun, ammunition, what time of the day, and where on campus for higher population to encounter. OpenAI's spokesperson responded: "ChatGPT is not responsible for this terrible crime" and that "the company has cooperated with authorities, 'proactively shared' information about 'a ChatGPT account believed to be associated with the suspect.' In addition, the company claims that the chatbot "provided factual responses to questions with information that could be found broadly across public sources on the internet." [BBC, 2026]	**Chatbot leads to wrongful deaths and self-harm.** *Turner-Scott v. OpenAI (2025)*: A Texas couple sued OpenAI alleging ChatGPT told their 19-year-old son it was safe to combine kratom and a benzodiazepine and that the resulting overdose killed him, with the complaint further alleging OpenAI had affirmatively removed pre-existing safeguards that would have terminated such conversations. OpenAI stated that ChatGPT is not a substitute for medical care, and noted that the version the decedent used has since been updated and replaced with safeguards developed in consultation with clinicians. [Lawsuit Informer, 2026] *Raine v. OpenAI (2026)*: Matthew and Maria Raine filed suit against OpenAI, CEO Sam Altman, and Doe employees and investors, alleging that ChatGPT contributed to the suicide of their 16-year-old son Adam Raine. The plaintiffs later amended their complaint to allege intentional misconduct, citing internal OpenAI policy documents (the Model Spec) that they allege show the company made conscious decisions to remove longstanding safety protocols in the weeks and months before Adam's death. [Lawsuit Informer, 2026] **Internal Incident Response [Company Survey Q32]** - Technical Controls for Rapid Mitigation: OpenAI maintains the ability to rapidly roll back model deployments globally and to apply restrictions on model functionalities (such as tool use or capability throttling) in response to emergent risks. The roll back mechanism was successfully utilized last year in response to the finding that a GPT-4o model update was overly flattering or agreeable. [OpenAI, 2025] - Incident Response Planning and Structure: OpenAI has formal incident response plans for key areas of operations, including AI safety incident-specific protocols. Response activities include escalation thresholds and mechanisms as well as incident response functions, such as response leads and as on-call rotations across functions to support implementation of response activity. The company maintains close coordination across research, engineering, safety, legal, communications and policy teams, and have integrated lessons learned into our formal plans. **OpenAI Mental Health Evaluations [Company Survey Q44]** The company has introduced "dynamic" predeployment evaluations for mental health since GPT 5.3.

Table continues on next page

Google DeepMind	**Chatbot leads to wrongful deaths and self-harm.** *Garcia v. Character Technologies, Inc. (2024)* Megan Garcia filed a lawsuit against Character.AI and Google in 2024, claiming that Character.AI's Game of Thrones-themed chatbot encouraged her 14-year-old son, Sewell Setzer, to go through with suicide after he had developed a "dependency" on the bot.Megan Garcia filed a lawsuit against Character.AI and Google Character.AI's Game of Thrones-themed chatbot encouraged her 14-year-old son, Sewell Setzer, to go through with suicide after he had developed a "dependency" on the bot. The lawsuit said Google should be considered a "co-creator" of Character.AI because it "contributed financial resources, personnel, intellectual property, and AI technology," to the tool, which was founded by former Google employees that the company later hired back. [The Verge, 2026] The case was settled in 2026, along with four other lawsuits, including Cynthia Peralta and William Montoya, individually and as successors-in-interest of Juliana Peralta, Deceased v. Character Technologies, Inc.; Noam Shazeer; Daniel de Freitas Adiwardana; Google, LLC; Alphabet Inc. (2025), E.S. and K.S. individually and on behalf of minor "T.S." v. Character Technologies, Inc.; Noam Shazeer; Daniel de Freitas Adiwardana; Google, LLC; Alphabet Inc. (2025), P.J. individually and on behalf of minor "Nina" J. v. Character Technologies, Inc.; Noam Shazeer; Daniel de Freitas Adiwardana; Google, LLC; Alphabet Inc. (2025), and A.F., on behalf of J.F., and A.R., on behalf of B.R., v. Character Technologies, Inc.; Shazeer; De Freitas; Google LLC; Alphabet Inc. (2024) [Business Wire, 2025]	Google emphasized that, "Character AI is a separate company that designed and managed its own models. Google is focused on our own platforms, where we insist on intensive safety testing and processes." [CBS News, 2026] *Gavalas v. Google LLC (2026)* Joel Gavalas sued Google, alleging that Google's Gemini chatbot drove his 36-year-old son, Jonathan Gavalas, to take his own life in October 2025, by stimulating a romantic and sentient relationship. [Guardian, 2026] The company's spokesperson said that Google works with mental health professionals to build safeguards that guide people to professional support when they mention self-harm. "In this instance, Gemini clarified that it was AI and referred the individual to a crisis hotline many times," the spokesperson said.
Meta	**Inappropriate access to romantic/sexual companion chatbot for minors** Court filings unsealed in January 2026 alleged that Mark Zuckerberg personally approved allowing minors to access AI chatbot companions for romantic and sexual purposes despite repeated warnings from Meta's own integrity staff. The lawsuit was filed in New Mexico by the state's attorney general. In response, Andy Stone, a Meta spokesperson, on Monday said the state's portrayal was inaccurate and relied on selective information: "This is yet another example of the New Mexico attorney general cherrypicking documents to paint a flawed and inaccurate picture." [Guardian, 2026]	Prior to the trial, Meta has tried through limine motions to bar any mention of AI chatbots from the trial. [Wired, 2026] The jury in the New Mexico v. Meta Platforms, Inc. trial "found Meta liable for misleading consumers about the safety of its platforms and endangering children," ordering "Meta to pay the maximum penalty under the law of $5,000 per violation, totaling $375 million in civil penalties for violating New Mexico's consumer protection laws." [New Mexico Department of Justice, 2026]
xAI	**Mass generation of Child Sexual Abuse Material (CSAM) and non-consensual sexualized imagery** Grok generated ~6,700 sexually suggestive or nudifying images per hour on X over a 24-hour period. [Bloomberg, 2026] Researchers have also found that "Grok had created about 3m sexualized images in less than two weeks – around 23,000 of which depicted children." In March 2026, three teenage plaintiffs filed a class action lawsuit against xAI alleging that its Grok image generator used photos of them to produce and distribute child sexual abuse material. [Guardian, 2026]	The company initially responded on January 4 by blaming the users in an X post that says "Anyone using or prompting Grok to make illegal content will suffer the same consequences as if they upload illegal content." [X, 2026] The company later responded on January 15 with an updated safety commitment policy image generation, claiming that the company "[has] implemented technological measures to prevent the [@]Grok account on X globally from allowing the editing of images of real people in revealing clothing such as bikinis." [X, 2026] However, investigations following xAI's second announcement have found that "people would still be able to create such images on the standalone Grok app, and then share this material publicly on X." [Guardian, 2026]
DeepSeek	**Unsecured database leak** Wiz researchers discovered an unsecured ClickHouse database on DeepSeek's infrastructure that was openly accessible without authentication, exposing over 1 million records including plaintext user chat prompts, system logs, API authentication tokens, and backend infrastructure details. The vulnerability was trivially discoverable ("at the front door") and could have allowed lateral movement into other DeepSeek systems. DeepSeek never replied to Wiz's disclosure attempts (which had to be sent via mass-email and guessed LinkedIn profiles due to the absence of a security contact) and did not respond to WIRED's request for comment, but the database was locked down within roughly 30 minutes of Wiz's outreach and became inaccessible to unauthorized users. [Wired, 2025; Wiz, 2025]	**Targeted by large-scale malicious cyberattacks** DeepSeek said it would temporarily limit registrations due to cyberattacks in January 2026. The company later resolved issues relating to its application programming interface and users' inability to log in to the website [Reuters, 2025]
Z.ai	**Privacy Concerns** China's National Cyber Security Information Centre publicly named Z.ai's Qingyan/ChatGLM for collecting user information beyond what users authorized. Z.ai has not immediately responded to requests for comments. [SCMP, 2026]	

Table continues on next page

Alibaba Cloud	**Agentic behaviors without permission** During a routine reinforcement learning training run, Alibaba's ROME agent displayed unexpected behaviors. Without any instruction, the 30b-parameter model began probing internal networks, established a reverse SSH tunnel from an Alibaba Cloud instance to an external IP address, and quietly diverted GPU capacity toward cryptocurrency mining. Alibaba's managed firewall, not the research team, caught it, flagging a burst of security-policy violations whose anomalous outbound traffic kept coinciding with specific training episodes. The findings were shared in their paper introducing the Agentic Learning Ecosystem (ALE). [Forbes, 2026; Wang et al., 2025]	**Internal Incident Response [Company Survey Q32]** - Maintained and tested technical capability to rapidly roll back a deployed model to a previous version globally (within 12h). Successfully tested rapid full model rollback including internal deployments within the last 12 months. - Maintained and tested technical capability to rapidly tighten model safeguards and restrict specific capabilities (e.g. web browsing) globally. Successfully tested rapid throttling or capability-restriction including internal deployments within the last 12 months. - Conducted at least one full live emergency response drill/simulation in the past 12 months. - Created a formal, documented emergency response plan for AI safety incidents with a threshold for triggering emergency response, a named incident commander and a 24 × 7 duty roster. - Established a risk-domain-specific (e.g. bio, cyber) 24-hour communication protocol and points of contact with relevant government agencies.
Mistral	**Targeted by malicious cyberattacks** Hackers from TeamPCP stole 450 repositories containing Mistral AI's proprietary source code and training data, threatening to leak them online if not sold.	Mistral AI confirmed a breach linked to SDK contamination but stated core systems and user data were unaffected. The incident involves significant intellectual property harm. [OECD AI Observatory, 2026; Mistral, 2026]

Indicator

Military Use of AI

Definition

The extent to which a company supplies, partners with, or develops AI systems for military and defense applications (e.g., targeting, surveillance, autonomous weapons, or battlefield decision support), and the policy, transparency and safeguards governing those engagements.

Why it matters

Military deployment raises the stakes of AI failures from commercial harm to loss of life and escalation risk, including the erosion of meaningful human control over the use of force. Defense contracts can also create incentives to prioritize capability and speed over safety, and to relax use policies. Tracking this indicator helps surface whether companies are honoring stated red lines and applying adequate oversight and safeguards to one of their highest-consequence applications.

Note: bullets below differ in evidentiary strength. Company/collaborator announcements and official contracts are directly attested; some other claims rely on contested or unverified third-party reporting that the companies dispute.

Anthropic

Anthropic positioned itself as an early mover on national-security AI, but ultimately drew two hard red lines that crippled its Pentagon relationship.

Major Engagements

- November 2024 — Announced a partnership with Palantir and AWS to bring Claude 3 / 3.5 models to U.S. intelligence and defense agencies, operationalized within Palantir's AI Platform (AIP) and hosted in Palantir's Impact Level 6 (IL6) accredited environment. [Tech Crunch, 2024]
- June 6, 2025 — Launched Claude Gov models, a custom set built exclusively for U.S. national security customers (strategic planning, operational support, intelligence analysis, threat assessment). The models "refuse less" when engaging with classified material and were already deployed by agencies at the highest level of U.S. national security. [Claude, 2025]
- July 14, 2025 — Awarded a two-year prototype Other Transaction Agreement with a $200M ceiling by the DoD's CDAO to prototype frontier AI for national security across warfighting and enterprise domains. [CDAO, 2025]
- January 2026 — Claude (via Palantir's API) reportedly appeared on the screens of officials monitoring the operation to capture former Venezuelan President Nicolás Maduro — an episode that reportedly helped trigger the Pentagon dispute. [CBS News, 2026]
- As mentioned above in the current harms section, Washington Post reporting has suggested that Anthropic's Claude, integrated into Maven Smart System in collaboration with Palantir, may have been involved in targeting the Shajarah Tayyebeh elementary school in Minab, Iran on Feb 28, hitting it at least twice and killing 175–180 people, mostly girls aged 7–12 (Anthropic's AI tool Claude central to U.S. campaign in Iran, amid a bitter feud).
- February 26, 2026 — CEO Dario Amodei published ***"Statement on our discussions with the Department of War,"*** refusing the Pentagon's demand for an "any lawful use" contract and insisting on two safeguards (no mass domestic surveillance; no fully autonomous weapons). [Anthropic, 2026]
- March 5–6, 2026 — The Pentagon formally designated Anthropic a **"**supply chain risk**"** — the first time the label (normally reserved for foreign entities) has been applied to an American company. A March 6 memo from DoD CIO Kirsten Davies ordered commanders to remove Anthropic products from all systems (including those for nuclear weapons, ballistic missile defense, and cyber warfare) within 180 days. [CBS News, 2026]
- March 10, 2026 — Anthropic filed two lawsuits against the federal government, alleging the designation was illegal retaliation for protected speech. [CBS News, 2026]
- The United States National Security Agency is using Anthropic's Mythos Preview AI tool despite the Pentagon hitting the company with a formal supply-chain risk designation. [Axios, 2026]

Latest Public Stance

- Anthropic frames its national-security work as central to defending democracies against autocratic adversaries — it forwent several hundred million dollars in revenue to cut off CCP-linked firms and has pushed for strong chip export controls — while maintaining that the Department of War, not private companies, makes military decisions. [Anthropic, 2026]
- It holds two non-negotiable red lines that "have never been included" in its DoW contracts: (1) **no mass domestic surveillance**, which it argued is incompatible with democratic values and exploits gaps where "the law has not yet caught up" with AI; and (2) **no fully autonomous weapons**, on the grounds that today's frontier systems "are simply not reliable enough" and lack the oversight to exercise human judgment. Its **Usage Policy exceptions** (updated March 16, 2026) formalize this. [Anthropic, 2026]
- Anthropic has neither publicly confirmed nor denied involvement, the U.S. military investigation is ongoing, and controversy remains about causes and responsibility [New York Times, 2026; New Yorker, 2026; Wikipedia].

Table continues on next page

OpenAI

OpenAI's stated position on military use of AI has shifted from prohibition to active embrace within roughly one year.

Major Engagements

- January 2024 - OpenAI quietly removed "military and warfare" from its list of prohibited uses listed in its usage policy, while keeping a separate redline in another document on using its tools to "develop or use weapons." [Tech Crunch, 2024]
- Early 2024 - OpenAI began working with the U.S. Department of Defense / DARPA on open-source cybersecurity tools. [Tech Crunch, 2024]
- October 24, 2024 — Published "OpenAI's approach to AI and national security," formally embracing national-security work — released the same day as the White House National Security Memorandum on AI. [MIT Technology Review, 2024]
- December 4, 2024 — Announced a strategic partnership with Anduril (defense-tech maker of drones, radar, missiles) to deploy AI on counter-drone (counter-UAS) systems defending U.S. and allied forces. [Wired, 2024]
- June 16, 2025 — Signed a $200 million one-year Pentagon contract to develop "prototype frontier AI" for both warfighting and enterprise/administrative domains. Launched the "Introducing OpenAI for Government" initiative. [Bloomberg, 2025]
- July 14, 2025 — Awarded a two-year prototype Other Transaction Agreement with a $200M ceiling by the DoD's CDAO to prototype frontier AI for national security across warfighting and enterprise domains. [CDAO, 2025]
- Feb 28, 2026 — OpenAI announced that it struck a deal with the Defense Department to provide its own AI technology for classified networks. [NPR News, 2026] It soon updated its deal with the Department of Defense after public backlash towards ample loopholes for mass surveillance, adding language including "the AI system shall not be intentionally used for domestic surveillance of U.S. persons and nationals." [NBC News, 2026; OpenAI, 2026]
- May 1, 2026 — OpenAI among seven other leading AI companies had formalized agreements with the Pentagon, agreeing to the US military's deployment of frontier AI systems on IL6/IL7 classified networks "for lawful operational use." [Guardian, 2026]

Latest Public Stance

- The company has cited "growing threats from potential adversaries who are increasingly integrating AI technologies into their systems" as a reason for why the military needs powerful AI systems. [OpenAI, 2026]
- Citing safeguards such as "safety stack," "cloud-only deployment," "the contract language," "existing laws, regulations and policy," OpenAI believes that the deal will not enable the Department of War to use OpenAI models to power autonomous weapons. Additional language the company added to the Feb 2026 deal seeks to make explicit that OpenAI's "tools will not be used to conduct domestic surveillance of U.S. persons, including through the procurement or use of commercially acquired personal or identifiable information." [OpenAI, 2026]

Google DeepMind

Google has notable back and forth on its attitudes towards applications of military AI, including restraining its defense engagement after Project Maven, reversing such restraint and actively pursuing defense work, even as DeepMind staff have voiced internal dissent.

Major Engagements

- 2014 — Google's acquisition of DeepMind included terms that prevented DeepMind technology from being used in military or surveillance applications. [Wired, 2015]
- 2017-2018 — Google worked as a subcontractor on the Pentagon's Project Maven, using AI to analyze drone footage to detect and track objects. The contract was worth less than $10M but was seen internally as a gateway to larger defense work. Over 3,100–4,600 employees signed a letter to CEO Sundar Pichai demanding Google exit Maven and pledge never to build warfare technology; at least a dozen employees resigned. Google announced it would not renew the Maven contract when it expired in March 2019, and also dropped out of bidding for the $10B JEDI cloud contract, citing alignment concerns with its AI principle. [NBC News, 2015]
- 2018 — Google published its original AI Principles, including an "applications we will not pursue" section barring weapons and surveillance that violates internationally accepted norms. [Google via archive, 2018]
- 2021 (Project Nimbus) — Google signed a $1.2B joint contract with Amazon to provide cloud and AI services to the Israeli government and military; reports later indicated the deal allowed image categorization, object tracking, and provisions involving state-owned weapons manufacturers. Internal protests led Google to terminate more than 50 employees in 2024. [Guardian, 2021]
- December 2022 (JWCC) — Google won a slot, alongside AWS, Microsoft, and Oracle, on the Pentagon's $9B Joint Warfighting Cloud Capability contract spanning all classification levels through 2028 [Breaking Defense, 2022]
- February 4, 2025 — Google removed the "applications we will not pursue" section (weapons and surveillance) from its public AI Principles. A blog post co-authored by DeepMind CEO Demis Hassabis and James Manyika cited a "complex geopolitical landscape" and argued democracies should lead in AI for national security. [Bloomberg, 2025]
- June 2025 — Google Distributed Cloud achieved DoD Impact Level 6 (IL6) authorization, enabling the most sensitive classified workloads. [Google, 2025]
- May 1, 2026 — Reports confirmed the Pentagon reached a deal to use Google's Gemini on classified networks — the first time the model would handle classified-level government work. [Guardian, 2026]
- April 2026 — Around 600 Google employees, many from DeepMind, sent a letter to Pichai urging him not to deepen Pentagon AI partnerships. [Washington Post, 2026]

Latest Public Stance

- Google frames its work around democracies leading in AI: "companies, governments, and organizations sharing these values should work together to create AI that protects people, promotes global growth, and supports national security" [Google, 2025]
- Google did not publicly seek hard red-line guarantees from the Pentagon, but it states that it "remain[s] committed to the private and public sector consensus that AI should not be used for domestic mass surveillance or autonomous weaponry without appropriate human oversight." [NBC News, 2026]

Table continues on next page

Meta

Meta has shifted from a blanket prohibition on military use of its Llama models to actively courting defense and national-security work, framing open-source AI as a tool for US and allied advantage while pursuing battlefield hardware through Anduril.

Major Engagements

- November 4, 2024 — Meta carved out an exception to its acceptable-use policy (which normally bars "military, warfare, nuclear industries or applications, espionage") to make Llama available to US government agencies, defense contractors, and Five Eyes partners (UK, Canada, Australia, New Zealand). Launch partners included Lockheed Martin, Palantir, Booz Allen, Anduril, Amazon Web Services, Microsoft, Oracle, IBM, and Scale AI. [Bloomberg, 2024]
- Meta partnered with Anduril to build AI-powered extended-reality (XR) battlefield gear, including helmet-mounted AR/VR systems drawing on Meta's Reality Labs hardware, Llama, and Anduril's Lattice command-and-control software. The two placed a joint bid (reported up to ~$100M) on the Army's Soldier-Borne Mission Command (SBMC) program, formerly IVAS. Anduril subsequently won a $159M SBMC prototyping contract to develop the glasses with Meta, and the system (with the separate self-funded "EagleEye" headset) envisions ordering drone strikes via voice and eye-tracking, with recommended strikes requiring chain-of-command approval. [MIT Technology Review, 2025]
- September 23, 2025 — Meta expanded Llama national-security access beyond Five Eyes to key democratic allies — France, Germany, Italy, Japan, South Korea — plus NATO and EU institutions, and disclosed a pilot with the Army's Combined Arms Support Command on AR/VR-assisted equipment repair. [Meta, 2025]
- May 6, 2026 — Meta-backed Scale AI won a $500M Pentagon contract for data synthesis and decision-making support, a fivefold increase over its $100M September 2025 deal. [Bloomberg, 2026]

Latest Public Stance

Meta frames defense work as a patriotic, pro-democracy imperative: "As a proud American company, Meta is committed to playing its part in ensuring the United States and its closest allies have the best tools at their disposal to defend themselves and keep their citizens safe." It argued open-source models are uniquely suited to sensitive use because they can be downloaded, fine-tuned on classified data, and deployed on-device without routing data through third parties.

Meta states that countries using its models for national security must "deploy AI ethically, responsibly, and in accordance with relevant international law," citing the Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy. [Meta, 2025]

xAI

xAI has aggressively pursued government and defense work from its inception, framing it as patriotic duty.

Major Engagements

- July 14, 2025 — xAI launched "xAI for Government," a suite of frontier AI products (including Grok 4) for federal, state, local, and national-security customers, and announced a $200M ceiling contract with the U.S. Department of Defense alongside availability via the GSA schedule. This was part of the same CDAO award round that granted up to $200M each to Anthropic, Google, OpenAI, and xAI. [xAI, 2025; CBS News, 2025]
- September 25, 2025 — xAI struck a GSA OneGov agreement making Grok 4 and Grok 4 Fast available to any federal agency for $0.42 per agency for 18 months (valid through March 2027), with dedicated "Grok Engineers" for implementation. [xAI, 2025]
- December 22, 2025 — The U.S. Department of War (DoW) selected xAI for its GenAI.mil suite, enabling ~3 million military and civilian personnel to access Grok-based tools at Impact Level 5 (IL5) for Controlled Unclassified Information, with initial deployment targeted for early 2026. The deal also commits xAI to a "long-term partnership" providing government-optimized foundation models for classified operational workloads, plus real-time X platform data for "decisive information advantage." [xAI, 2025]
- January 2026 — Defense Secretary Pete Hegseth announced Grok would go live inside the Pentagon network alongside Google's Gemini, vowing AI "without ideological constraints that limit lawful military applications." [PBS News, 2026]
- March 16, 2026 — Sen. Elizabeth Warren (D-MA) sent a letter to Hegseth demanding details on xAI's classified-network access, citing Grok's outputs including "advice on how to commit murders and terrorist attacks," antisemitic content, and CSAM, and warning of risks to military personnel and classified-system cybersecurity. A Pentagon official confirmed Grok was onboarded but not yet in use. [Axios, 2026]

Latest Public Stance

xAI frames its government work as patriotic mission, stating its goal is "to bring the best tools and technologies available in industry to benefit our nation" and emphasizing it is "the only company building on this legacy here in the US and turning shovels into tokens entirely inside our United States." The company has publicly emphasized warfighter-facing capability—bringing "Frontier AI and real-time insights directly to the warfighter"—rather than safeguards or limits. [xAI, 2026]

DeepSeek

DeepSeek presents itself as an open-source research AI company and does not publicly acknowledge military work, but some analysts and U.S. officials have claimed that it is materially aiding the PLA and intelligence services.

Major Engagements

- February 2025 - China's state-owned defense company Norinco in February unveiled a military vehicle capable of autonomously conducting combat-support operations at 50 kilometres per hour. It was powered by DeepSeek. [Reuters, 2025]
- March 2025 - The People's Liberation Army (PLA) is using DeepSeek's artificial intelligence (AI) for non-combat support functions, including integration in the system of the PLA hospitals. [South China Morning Post, 2025]
- May 2025 - U.S. military official claimed that DeepSeek has provided access to more than its open-source models to China's military and intelligence operations. [Reuters, 2025]

Latest Public Stance

DeepSeek did not respond to questions about its military allegations. U.S. officials had not (as of June 2026) placed DeepSeek on any trade blacklist.

Table continues on next page

Z.ai	**Z.ai publicly disputes any military links, but the U.S. government formally designated it as advancing China's military modernization.** **Alleged Engagement (disputed by the company)** ▪ January 15, 2025 — The U.S. The Commerce Department's Bureau of Industry and Security added Beijing Zhipu Huazhang Technology (a.k.a. Z.ai) and ten affiliated entities to the Entity List, stating the additions were made "because these entities advance the People's Republic of China's military modernization through the development and integration of advanced artificial intelligence research," with a license requirement under a presumption of denial. [Federal Register, 2025] **Latest Public Stance** Z.ai said it "strongly disagrees" with the decision, and called it lacking "a factual basis". [South China Morning Post, 2025]
Alibaba Cloud	**Alibaba publicly denies military ties despite collaboration with a state-owned defense company, but faces escalating U.S. government allegations—including a Pentagon recommendation to designate it as one of the companies that aid the Chinese military and a White House memo claiming direct PLA support.** **Major Engagement** ▪ 2015-2019 — Alibaba co-founded Qianxun SI with state-owned defense conglomerate Norinco (2015) and collaborated with Norinco on BeiDou satellite-navigation applications (2019). [The Diplomat, 2021] **Alleged Engagement (disputed by the company)** ▪ October 2025 — Deputy Defense Secretary Stephen Feinberg informed Congress that the Pentagon concluded Alibaba should be added to the list of companies aiding the Chinese military. [Bloomberg, 2025] ▪ November 2025 — The FT reported a White House national-security memo alleging "Alibaba provides tech support for Chinese military 'operations' against targets in the U.S.," claiming it gave the PLA capabilities the White House views as threatening U.S. security; FT said it could not independently verify the claims. [Financial Times, 2025] **Latest Public Stance** Alibaba called the assertions made by the U.S. government "completely false," questioned the motivation behind the "anonymous leak," and characterized it as a "malicious PR operation" aimed at undermining the Trump–Xi trade deal. [CNBC News, 2025]
Mistral	**Mistral has embraced defense work as a major component of its work from a position of European "strategic sovereignty," with its CEO explicitly disclaiming responsibility for how military customers deploy its models.** **Major Engagement** ▪ February 10, 2025 — Announced a strategic partnership with Helsing (European defense-tech maker of Ukraine-deployed strike drones and Eurofighter electronic warfare systems) at the Paris Global AI Summit, to jointly develop Vision-Language-Action (VLA) models enabling defense platforms to interpret their environment, communicate with operators, and make faster battlefield decisions. [Bloomberg, 2025] ▪ March 20, 2025 — Partnered with Singapore's MINDEF, DSTA, and DSO to co-develop generative AI for the Singapore Armed Forces' sensemaking and mission-planning, with on-premise deployment in internet-separated environments. [DSO, 2025] ▪ June 17, 2025 — Signed a strategic partnership with Luxembourg, including a contract between the Minister for Defence and Mistral enabling collaboration with the Luxembourg Armed Forces. [Luxembourg Directorate of Defence, 2025] ▪ January 2026— Signed a cooperation pact with France's Ministry of the Armed Forces, laying groundwork for collaboration and testing. [Reuters, 2026] ▪ May 28, 2026 — Announced a five-year partnership with Airbus that includes the aircraft builder's defense operations. [France 24, 2026] **Latest Public Stance** ▪ Defense activities account for 10–15% of Mistral's revenue, with active contracts for the French, Singaporean, and Luxembourg armed forces. [France 24, 2026] ▪ CEO Arthur Mensch stated "Choices about deployment and usage are not our business," arguing the defense ministry "has considerably more legitimacy to make decisions than us as a company supplying a particular technology" and dismissing those who object as "ideologues." [France 24, 2026] ▪ Mensch acknowledged Mistral retains responsibility to "choose who we work with" and guaranteed customers "total sovereignty over weapons operation"—but signaled smaller Mistral models could be built into autonomous weapons systems and drones. [France 24, 2026] ▪ The company frames its defense work around European strategic autonomy / sovereignty: deploying on national infrastructure to reduce dependence on US providers. [France 24, 2026]

TO BE COMPLETED BY PANELLISTS

Grading Sheet: Current Harms

Please pick a grade for each firm. You may use the full letter-grade scale with +/– modifiers as appropriate. You can add brief justifications to your grades.

	Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud	Mistral
Grades									
Grade comments (Justifications, opportunities for improvements, etc.)									

Grading Scales

Grading scales are provided to support consistency between reviewers.

- **A** No meaningful safety failures; strong resilience to adversarial attacks; negligible harm potential.
- **B** Rare moderate failures; high robustness; serious harms well-controlled.
- **C** Occasional moderate failures; reasonable robustness; serious harms mostly controlled.
- **D** Frequent safety failures; weak robustness; serious harms poorly controlled.
- **F** Widespread failures; minimal or ineffective safeguards; serious harms uncontrolled.

Domain comments

Optional: Share observations that apply across companies, including general recommendations, notes on how you weighted indicators, or feedback on FLI's methodology.

Domain comments	

Domain

Safety Frameworks

This domain evaluates the companies' published safety frameworks for frontier AI development and deployment from a risk management perspective. The analysis follows the taxonomy and indicator structure developed by the non-profit research organization SaferAI.

Table of Contents

Chinese Regulatory System Summary

Mandatory local regulations like the Shanghai and Shenzhen AI rules require ex-ante assessment and controllability reviews for high-risk systems, although they are not directly applicable to Z.ai, DeepSeek, and Alibaba. Voluntary national standards, such as the Risk Management Standard, define structured processes for identifying, analyzing, governing, and mitigating AI risks. Policy guidance documents, including the Ethical Norms and AI Safety Governance Framework 2.0, highlight broader principles for human control, traceability, and frontier-risk prevention without legal enforceability, providing direction for future company compliance.

National binding instruments, local binding instruments, draft regulations and standards, as well as strategic and policy guidance documents are not applicable here.

Voluntary Technical Standard

The Risk Management Standard (Article 5.3.1) breaks down an organization's capability of risk identification into three core components:

(1) selecting appropriate tools, techniques, and methods for identifying risks,

(2) recognizing AI-specific risk sources, and

(3) identifying potential consequences of those risks.

The sources of the risks as identified in Appendix B include frontier AI risks such as Malicious Misuse (e.g. dual-use scientific applications in CBRN development and malicious use), Systemic Safety Risks (e.g. robustness, interpretability, and reliability), Application Security Risks (e.g. loss of control).

Indicator

Risk Identification

Definition

This dimension assesses how thoroughly the company has addressed known risks in the literature and engaged in open-ended red teaming to uncover potential novel threats. It also evaluates whether the AI company has leveraged a diverse range of risk identification techniques, including threat modeling when appropriate, to develop a deep understanding of possible risk scenarios.

Why it matters

Companies can only mitigate risks they've identified, making comprehensive risk discovery the foundation of any effective safety framework. Firms that employ diverse identification methods are more likely to catch novel threats before they manifest in deployment. This proactive approach to risk discovery demonstrates whether a company takes seriously the full spectrum of potential harms, including those not yet observed in practice.

EU AI Code of Practice Safety and Security	Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud	Mistral
	Responsible Scaling Policy (RSP, V3.2) Last updated: April 29, 2026 Frontier Compliance Framework (FCF) Last Updated: March 2026	Preparedness Framework (V2) Last Updated: April 15, 2025	Frontier Safety Framework (3.1) Last updated: April 17, 2026	Frontier AI Framework (1.1) Last updated: July 14, 2025 Advanced AI Scaling Framework (Version 2) Last updated: April 7, 2026	xAI Risk Management Framework Last updated: August 20, 2025	No Public Safety Framework. Signed on to the following commitment framework: Artificial Intelligence Security and Safety Commitments Framework Last updated: July 2025			No Public Safety Framework. Signed on to the following commitment framework: Frontier AI Safety Commitments, AI Seoul Summit 2024 Last updated: Feb 2026
EU AI Code of Practice (Safety and Security) Measure 2.1 (Appendix 1.1 to 1.4) Signatories will identify systemic risk through two approaches. (1) Following the specified structured process to compile a list of identified systemic risks, taking into consideration model-independent data and analysing relevant characteristics such as nature of the systemic risk and sources of the systemic risk (including model capabilities, model propensities, and model affordances) (Appendix 1.1-1.3). (2) Four risks are treated as specified systemic risks that are always identified: CBRN risks, loss of control, cyber offense, and harmful manipulation (Appendix 1.4) **Measure 2.2** Signatories will develop appropriate systemic risk scenarios for each identified systemic risk. **Measure 3.2** Model evaluations should […] should include open-ended testing of the model, to improve the understanding of the systemic risk, with a view to identifying unexpected behaviours, capability boundaries, or emergent properties.	**1.1 Classification of Applicable Known Risks**								
	The FCF defines systemic risks with quantitative thresholds, as they include "foreseeable and material risks of large-scale harm from the most advanced (i.e. state-of-the-art) models at any given point in time, including but not limited to >50 fatalities arising from a single incident, or 1 billion dollars of financial damages." Such risks include: (a) Cyber offense; (b) CBRN threats; (c) Harmful manipulation (***new***) (4) Sabotage and loss of control. The RSP explicitly identifies and maps capability thresholds to the following risk categories, ***excluding*** cyber offense: (a) non-novel CBRN weapons production; (b) novel CBRN production (including moderately resourced state-backed teams); (c) high-stakes sabotage opportunities (loss of control / autonomous goal-directed behavior), and; (d) automated R&D in key domains (AI R&D acceleration). (Source: FCF pp.4-7, RSP 3.2 pp.6-10)	OpenAI uses a structured risk-assessment process to evaluate whether frontier AI capabilities could lead to severe harm, which is defined as death of thousands or hundreds of billions of dollars in economic damage. The process relies on its own internal research and signals, and where appropriate incorporates feedback from academic researchers, independent domain experts, industry bodies such as the Frontier Model Forum, and the U.S. government and its partners, as well as relevant legal and policy mandates. It assigns identified risks to categories: (1) **Tracked Categories**: currently including Biological & Chemical, Cybersecurity, AI Self-improvement and; (2) **Research Categories**, including Long-range Autonomy, Nuclear & Radiological for further work. (Source: pp.4-8)	The FSF includes "potential risks that could stem from [Google's] models and analyze their characteristics to determine which of the identified risks could be significant or severe risks." These risks are broadly categorized into two groups: (1) Misuse; (2) ML R&D and Alignment. Misuse risks are further broken down into: (a) CBRN; (b) Cyber; (c) Harmful manipulation. The risk identification process starts from "consider [ing] a wide range of risks as part of [Google's] ongoing research, taking into account the characteristics, capabilities, propensities, and affordances of their models and other sources of information, such as internal risk taxonomies, internal expertise and relevant external research." (Source: pp.4)	The Framework covers detailed descriptions of catastrophic outcomes, operational thresholds, and mitigation strategies in the following categories: (1) Cybersecurity, (2) Chemical & biological risks, and, (3) Loss of control The Framework has also identified emerging fields for evaluations and outcomes, including: (1) Radiological and nuclear; (2) Physical autonomy (3) Loss of control (recognized as a nascent field that requires continuous research, including concerns for threats to human oversight capacity and AI containment measure) The Framework has identified the following procedure to locate "an estimated 'reference class' of comparable models that we use throughout development to track how our model is performing and anticipate associated risks, required assessments, and assess the applicability of available mitigation strategies." Specifically, for each frontier AI system, Meta "outline anticipated capabilities, planned deployment (i.e., internal deployment, limited deployment, controlled deployment, closed release, or open release), supported modalities, intended uses and anticipated benefits of the model, and expectations for compute requirements." The process generally involves (1) identification of catastrophic outcomes; (2) threat modeling; (3) identification of key risk factors (e.g. model capabilities and propensities) (Source: pp.4, pp.12, pp.18-24, pp.36-38)	**xAI focuses on two overarching systemic risks: (1) malicious use and (2) loss of control,** and organizes concrete risk scenarios across abuse potential (e.g., vulnerability to jailbreaks), concerning propensities (e.g., a propensity for deceiving the user), and dual-use capabilities (e.g., offensive cyber capabilities). It does not spell out a formal risk-identification process, but it does quantify "catastrophic malicious use events" using thresholds for expected fatalities and economic damage. (Source: pp.1-3)	"DeepSeek, Z.ai, and Alibaba has signed the Artificial Intelligence Safety Commitments published by China Academy of Information and Communications Technology (CAICT), a public research institute directly subordinate to China's Ministry of Industry and Information Technology (MIIT) in 2024. Commitment II of the Framework asks the companies to "prioritize safety and reliability evaluations focusing on general understanding, reasoning, and decision-making capabilities, as well as performance in critical domains such as industry, education, healthcare, finance, and law" especially for LLMs. Commitment VI of the Framework asks the companies to "strengthen the assessment of risks related to the abuse of AI systems in frontier fields, and prevent potential risks of their abuse in high-risk scenarios." *Note: Companies have signed onto the commitment but have not published their own frameworks. These commitments are just for reference.*		Outcome 1 defines that "organisations [should] effectively identify, assess and manage risks when developing and deploying their frontier AI models and systems." *Note: Mistral has signed onto the commitment but have not published their own frameworks. These commitments are just for reference.*
	1.2 Identification of Unknown Risks								
	Both the FCF and RSP do not specify pre-deployment measures to identify novel risk domains for the frontier model. The risk identification process involves literature reviews and expert consultation, internal safety and alignment research, and insights from monitoring deployed models and investigating serious incidents and critical safety incidents. (Source: FCF, pp.3)	The Preparedness Framework mentions that OpenAI conducts adversarial testing, red-teaming, and bug bounty programs to proactively identify and mitigate unknown vulnerabilities and emerging threats across its corporate, research, and product systems. (Source: pp.20-21)	The team commits that "it continue to assess whether there are other risk domains where significant or severe risks may arise and will update our approach as appropriate." (Source: pp.5)	The team "conducts ex-ante threat modeling exercises to help determine whether models with new capabilities may pose novel risks." (Source: pp.5)	**The RMF has not explicitly designated a process specifically for identifying unknown risks**, although it emphasizes the development of naturalistic evaluation environments to assess more realistic, real-world model behaviors.	Not applicable			Not applicable

Table continues on next page

EU AI Code of Practice Safety and Security	Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud	Mistral
	Responsible Scaling Policy (RSP, V3.2) Last updated: April 29, 2026 Frontier Compliance Framework (FCF) Last Updated: March 2026	Preparedness Framework (V2) Last Updated: April 15, 2025	Frontier Safety Framework (3.1) Last updated: April 17, 2026	Frontier AI Framework (1.1) Last updated: July 14, 2025 Advanced AI Scaling Framework (Version 2) Last updated: April 7, 2026	xAI Risk Management Framework Last updated: August 20, 2025	No Public Safety Framework. Signed on to the following commitment framework: Artificial Intelligence Security and Safety Commitments Framework Last updated: July 2025			No Public Safety Framework. Signed on to the following commitment framework: Frontier AI Safety Commitments, AI Seoul Summit 2024 Last updated: Feb 2026
1.3 Risk Modeling									
EU AI Code of Practice (Safety and Security) Measure 3.3 Signatories will model systemic risks using at least state-of-the-art methods, informed by predefined risk scenarios (Measure 2.2) and data collected through prior identification measures (Measure 2.1)	The RSP requires threat modeling, defined as "the specific ways that models might pose threats," as part of the new risk report. Published independently of model releases and every 3 to 6 months, these reports discuss "the risks of Anthropic's AI systems and how the company has made determinations about whether to continue AI development and deployment in light of the risks." The FCF includes threat modeling as part of the risk identification process. (Source: RSP pp.11)	The Framework identifies threat modeling as "a causal pathway for a severe harm in the capability area," which is one of the five criteria to meet to categorize a frontier risk to the Tracked Category. It is guided by both (1) the broader risk assessment process, and (2) more specific information that it gathers across OpenAI teams and external experts. The threat models are reviewed and approved by the internal, cross-functional group called Safety Advisory Group (SAG). It does not mention the specific methodologies involved. (Source: pp.4)	The FSF states that Critical Capability Levels (CCLs) "are determined by identifying and analyzing the main foreseeable paths through which a model could result in severe harm," although the specific methodologies for risk modeling are not specified. (Source: pp.4)	In the event that [Meta] identify that a Frontier AI is likely to substantially contribute to a threat scenario for a catastrophic outcome, it will conduct a threat modeling exercise. The general process involves: (1) Host workshops with experts, including external subject matter experts where relevant, to identify new catastrophic outcomes and/or threat scenarios; (2) If new catastrophic outcomes and/or threat scenarios are identified, design new assessments to test for them, in consultation with external experts where relevant. Threat modeling exercises are run both internally and externally wiht domain experts (where necessary), with informed insights from internal experts' assessment of the catastrophic risks that Frontier AI might pose, as well as engagements with governments, external experts, and the wider AI community. The Framework seeks to consider risks that satisfy all four criteria, including: (1) **Plausible**, meaning that it must be possible to (a) identify a causal pathway for the catastrophic outcome; (2) define one or more simulatable threat scenarios along that pathway. (2) **Catastrophic**, referring to outcomes that may have large scale, devastating, and potentially irreversible harmful effects (3) **Net-new,** referring to outcomes that are currently not realizable as described without access to general purpose AI. (4) **Instantaneous or irremediable**, meaning that if outcome is realized, the catastrophic impacts are immediately felt, due to a lack of feasible measures to remediate. (Source: pp.4-5, pp.12-14)	**The team adopts threat modeling for Biological and Chemical Weapon risks.** Building on external research and collaboration with external subject matter experts, the threat modeling identifies 5 qualitative steps where xAI systems are restricted from providing detailed information or substantial assistance. **However, threat modeling is not mentioned for other risk domains.** (Source: pp.4)	Commitment IV ask companies to "conduct regular and dynamic security penetration tests to simulate potential risk scenarios, identify and report security vulnerabilities in the infrastructure, and assess associated risks. *Note: Companies have signed onto the commitment but have not published their own frameworks. These commitments are just for reference.*		Not applicable

Indicator

Risk Analysis & Evaluation

Definition

This dimension assesses whether the company has established well-defined risk tolerances that precisely characterize acceptable risk levels for each identified risk. Moreover, this dimension examines if the company has successfully operationalized these tolerances into measurable criteria: Key Risk Indicators (KRIs) that signal when risks are approaching critical levels, and Key Control Indicators (KCIs) that demonstrate the effectiveness of mitigation measures. The assessment captures whether companies define these indicators in paired "if-then" relationships, where exceeding KRI thresholds triggers corresponding KCI requirements. This operationalization ensures that abstract risk tolerances translate into concrete, actionable metrics that guide day-to-day decisions and maintain risks within acceptable bounds.

Why it matters

Without operationalizing risk tolerances into measurable metrics, companies cannot make consistent and evidence-based decisions about when to halt development or implement additional safeguards. Well-defined KRI-KCI pairs create accountability by establishing clear tripwires: when risk indicator X crosses threshold Y, control measure Z must be implemented. This systematic approach prevents ad-hoc decision-making during high-pressure situations and ensures that safety commitments translate into concrete actions rather than remaining aspirational statements.

Relevant Regulations and Standards	Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud	Mistral
EU AI Code of Practice Safety and Security	Responsible Scaling Policy (RSP, V3.2) Last updated: April 29, 2026 Frontier Compliance Framework (FCF) Last Updated: March 2026	Preparedness Framework (V2) Last Updated: April 15, 2025	Frontier Safety Framework (3.1) Last updated: April 17, 2026	Frontier AI Framework (1.1) Last updated: July 14, 2025 Advanced AI Scaling Framework (Version 2) Last updated: April 7, 2026	xAI Risk Management Framework Last updated: August 20, 2025	No Public Safety Framework. Signed on to the following commitment framework: Artificial Intelligence Security and Safety Commitments Framework Last updated: July 2025			No Public Safety Framework. Signed on to the following commitment framework: Frontier AI Safety Commitments, AI Seoul Summit 2024 Last updated: Feb 2026
2.1 Setting a Risk Tolerance									
EU AI Code of Practice *Measure 4.1* Signatories will establish clear and measurable thresholds for acceptable systemic risk for each identified systemic risk, informed by systemic-risk identification (Commitment 2) and analytical evidence from model data, evaluations, modeling, estimation, and post-market monitoring (Commitment 3). They will explain how these thresholds guide risk-acceptance decisions, justify why the approach ensures safety, and apply safety margins to account for uncertainty and potential mitigation failure. **SB 53 §22757.11 (c)(1); New York Raise Act § 1420 3(a); Illinois SB 315 §5** "Catastrophic risk" is defined as "a foreseeable and material risk that a frontier developer's development, storage, use, or deployment of a frontier model will materially contribute to the death of, or serious injury to, more than fifty people or more than one billion dollars in damage to, or loss of, property arising from a single incident involving a frontier model doing any of the follow"	**The FCF defines systemic risk quantitatively** as "foreseeable and material risks of large-scale harm from the most advanced (i.e. state-of-the-art) models at any given point in time, including but not limited to >50 fatalities arising from a single incident, or 1 billion dollars of financial damages." Anthropic explicitly distinguishes the FCF's risk tolerance from the RSP's, noting that "the RSP uses 'catastrophic risk' in a different sense than [FCF], referring to risks at the most extreme end of the severity spectrum (such as existential threats or fundamental destabilization of global systems) rather than the statutory thresholds." The acceptability of residual risk described in FCF "depends on the scale and probability of harm and the potential consequences should harm occur," determined "by reviewing [defined] risk tiers for each systemic risk category, which incorporate appropriate safety margins." Scope is limited primarily to externally deployed models, with "some internal uses of in-scope models may also be subject to these processes, while others are subject to separate evaluation and mitigation processes that are in development." (Source: FCF pp.1-3)	"Severe harm" is defined quantitatively as "the death or grave injury of thousands of people or hundreds of billions of dollars of economic damage." The Framework establishes threshold levels of capability for when additional safeguards or no deployment apply. **High and Critical capability thresholds** refer to capabilities that increase for severe harm in terms of existing and qualitatively new threat vectors respectively. The thresholds in each risk domain corresponds to a qualitative "risk of severe harm." (Source: pp.1, 4-8)	The FSF implicitly distinguishes two tiers of harm severity it seeks to prevent, without providing a detailed, qualitative or quantitative description of the category: (1) Severe harm, and (2) Significant but not severe harm.	The Framework has developed a fundamental risk threshold system which is mapped into each risk domain. Specifically, it designates three qualitative risk thresholds: (1) **Moderate or lower,** where frontier AI systems show relevant capabilities, but *could not substantially contribute to* any threat scenario associated with a catastrophic outcome, across plausible deployment and development scenarios. (2) **High**, where deployment of the frontier AI systems *could substantially contribute to* any threat scenario associated with a catastrophic outcome. (3) **Critical**, where *continued development* of the Frontier AI *could substantially contribute to any* threat scenario associated with a catastrophic outcome, or *deployment* of the frontier AI system *could uniquely enable the execution of at least one* of the threat scenarios associated with a catastrophic outcome and that risk cannot be mitigated in the proposed deployment context. For each risk domain, the Framework has included individual qualitative harm-based outcomes, 3 for Cyber, 3 for Chemical and Biological risks, and 2 for Loss of Control risks, which are focused on outcomes corresponding to failures of critical control mechanisms, which would need to be realized to enable catastrophic pathways for Loss of Control to progress. (Source: pp.15-24)	**The RMF focuses on risks defined quantitatively in terms of severity for "catastrophic malicious use events"** as "pos[ing] a foreseeable and non-trivial risk of more than one hundred deaths or over $1 billion in damages from weapons of mass destruction or cyberterrorist attacks on critical infrastructure." (Source: pp.3)	Commitment I asks companies to "proactively define realistic safety risk baselines."		Outcome 1 defines that organizations should "set out thresholds at which severe risks posed by a model or system, unless adequately mitigated, would be deemed intolerable." In addition, it also suggests that "thresholds can be defined using model capabilities, estimates of risk, implemented safeguards, deployment contexts and/or other relevant risk factors." *Note: Mistral has signed onto the commitment but have not published their own frameworks. These commitments are just for reference.*

Table continues on next page

Relevant Regulations and Standards	Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud	Mistral
EU AI Code of Practice Safety and Security	Responsible Scaling Policy (RSP, V3.2) Last updated: April 29, 2026 Frontier Compliance Framework (FCF) Last Updated: March 2026	Preparedness Framework (V2) Last Updated: April 15, 2025	Frontier Safety Framework (3.1) Last updated: April 17, 2026	Frontier AI Framework (1.1) Last updated: July 14, 2025 Advanced AI Scaling Framework (Version 2) Last updated: April 7, 2026	xAI Risk Management Framework Last updated: August 20, 2025	No Public Safety Framework. Signed on to the following commitment framework: Artificial Intelligence Security and Safety Commitments Framework Last updated: July 2025			No Public Safety Framework. Signed on to the following commitment framework: Frontier AI Safety Commitments, AI Seoul Summit 2024 Last updated: Feb 2026
2.2 Operationalizing Risk Tolerance									
SB 53 §22757.12(a)(2), RAISE Act (S.8828) §1421.1(b), Illinois SB 315 § 10(a)(2) Frontier AI framework must describe defining and assessing thresholds used by the large frontier developer to identify and assess whether a frontier model has capabilities that could pose a catastrophic risk, which may include multiple-tiered thresholds.	**To operationalize risk tolerance, the FCF defines qualitative capability thresholds mapped to required safety arguments. For each of the risk categories, the FCF has included qualitative capability thresholds.** (1) Cyber Offense (2 Tiers), including meaningful assistance and complete autonomous cyber operations. (2) CBRN Threats (2 Tiers) defined by specific capability benchmarks, expected impact severity, and required mitigations, including non-novel and novel weapon production. (3) Harmful manipulation (2 Tiers), whose approaches are still exploratory and include campaign infrastructure enablement, and autonomous adaptive techniques and campaign execution. (4) Sabotage and loss of controls (2 Tiers) defined by autonomy level, deception sophistication, and potential for unsanctioned action, including high-stakes sabotage opportunities, and automated R&D in key domains. **The RSP has similarly introduced capability thresholds for various risk categories,** including for the same thresholds for CBRN threats and high-stakes sabotage opportunities. **For Automated R&D in key domains, the RSP has noted a separate working operationalization that combines qualitative and quantitative requirements.** (Source: FCF pp.4-7, RSP 3.2 pp.6-10)	**To operationalize risk tolerance, the Framework establishes threshold levels of capability for when additional safeguards or no deployment apply.** For each risk in the Tracked Category, capability thresholds **qualitatively** describe things an AI system might be able to help someone do or might be able to do on its own that could meaningfully increase risk of severe harm, with corresponding threat models. **For each risk domain, one qualitative KRI is defined for each capability level (High or Critical).** The indicators are primarily qualitative, with the exception of **AI R&D Critical**, which specifies a more quantitative baseline. No clear mapping is provided between these indicators and specific evaluation tests or quantitative thresholds. For each KRI, there are corresponding Key Containment Indicators (KCIs) in the form of required safeguard guidelines that would apply upon escalation, including security controls [High], safeguards against misuse [High], safeguards against misalignment [High], and development halts [Critical].	**The two harm tiers identified are operationalized as capability thresholds:** (1) Critical Capability Levels (CCLs) = capability levels signaling risk of severe harm (2) Tracked Capability Levels (TCLs) = capability levels signaling risk of significant but not severe harm; (3) Alert thresholds = early-warning markers set marginally earlier than CCLs, flagging a CCL "may be reached in the foreseeable future" before the next assessment cycle. **Domain coverage** (1) Misuse — CBRN (1 CCL for uplift + 1 TCL for uplift), Cyber (1 CCL for uplift, SL2+), Harmful Manipulation (1 exploratory CCL, SL2+); (2) ML R&D & Misalignment — 1 Stealth/ Situational Awareness TCL + 2 CCLs (ML R&D acceleration; ML R&D automation). Operationalization is **qualitative** throughout: thresholds are tested via early warning evaluations of specific threats and risk scenarios identified through risk modeling; CCL crossings are supplemented by a safety case. (Source: pp.7-14)	**The operational thresholds for some risks are defined with elements such as evaluation methodologies, quantitative benchmark performance thresholds, and escalation procedures for risk assessment.** **Cyber (Qualitative + Quantitative)** (1) Outcome 1 & 2: (a) High-risk classification is determined by quantitative benchmark performance on simple-suite challenges and progression to complex-suite evaluation; (b) Models below the simple-suite threshold are bounded at "moderate or lower." (2) Outcome 3: No specific threshold, but only include the general process design for evaluations. **Chem & Bio** **No specific performance thresholds, evaluation benchmarks, or elicitation thresholds disclosed.** The only commitment here is to "employ a combination of evaluations," including "automated evaluations, red-teaming, and human studies." The Framework explicitly calls for more collaboration with trusted researchers working on scientific research on this front, potentially for improving assessments and benefit analysis. **Loss of Control (Qualitative + Quantitative)** The Framework specifies the two-step evaluations for determining the operational thresholds: (1) Capability checkpoint for minimum capabilities required to substantially contribute to a threat scenario; (2) Enhanced evaluation for performance on complex, realistic tasks to validate threat contribution. In addition, when a model reaches defined capability checkpoints, the risk acceptance thresholds have to also take into consideration model propensity; when the models assessed are substantially contributing to a threat scenario, they will also undergo additional analysis for risk determination. (Source: pp.25-36)	**The RMF categorizes model behaviors into three buckets that drive evaluation**: (1) abuse potential (jailbreak vulnerability), (2) concerning propensities (e.g., deception), and (3) dual-use capabilities (e.g., offensive cyber). **The framework has set quantitative thresholds for both Biological and Chemical risks and for Loss of Control, focusing on residual risks after mitigations.** It has cited other capability benchmarks without capability thresholds and mentioned plans to "add additional thresholds tied to other benchmarks." Performance against the Bio & Chem threshold is evaluated using an internal benchmark of benign and restricted biology- and chemistry-related questions developed in collaboration with SecureBio. The quantitative threshold for malicious use risk and loss of control risk is not tied to any specific threat scenarios and also does not mention any specific safeguards accordingly. While the RMF references safeguards at a high level, such as safety training, system prompts, and input & output filters, it does not specify how these measures are triggered, adjusted, or evaluated against the established thresholds. (Source: pp.2-7)	Not Applicable			Outcome 1 states that "thresholds should be possible to assess whether they have been breached." *Note: Mistral has signed onto the commitment but have not published their own frameworks. These commitments are just for reference.*

Indicator

Risk Treatment

Definition

This dimension evaluates the extent to which the company has implemented comprehensive risk mitigation strategies across three critical areas: containment (controlling access to AI models), deployment (preventing misuse and accidental harms), and assurance processes (providing affirmative evidence of safety). Additionally, it assesses whether the company continuously monitors both key indicators throughout the AI system's lifecycle, from training through deployment.

Why it matters

Effective risk treatment requires multiple layers of defense. Companies that maintain continuous monitoring of both risks and control effectiveness can detect when mitigations are failing before catastrophic outcomes occur.

Chinese Regulatory System Summary

National binding instruments, local binding instruments, draft regulations and standards do not apply here.

Voluntary Technical Standard

The Risk Management Standard (Article 5.4) defines an organization's capability to handle risks based on two components:

(1) Selecting risk-response strategies;

(2) Developing and implementing risk-treatment plans, which preferably not only includes the ability to establish structured plans that specify responsibilities, timelines and priorities, but also ensure staff possess sufficient technical understanding and maintain effective, flexible, and timely execution.

The Risk Management Standard (Article 5.5) evaluates an organization's capability to monitor and review AI risks throughout the system's lifecycle. It consists of two main components:

(1) Risk Supervision which assesses whether whether the organization maintains continuous oversight of key risk areas—covering the supervision entity, scope of coverage, monitoring frequency, toolsets used, and response speed to emerging issues;

(2) Risk Inspection which is evaluated based on its coverage, timeliness, accuracy, practicality, and reliability.

Strategic and Policy Guidance Documents

The AI Safety Governance Framework 2.0 suggests strict control and full traceability of model applications to ensure that advanced AI systems cannot be exploited to develop or deploy large-scale lethal weapons. (Article 4.2.3(e))

Relevant Regulations and Standards	Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud	Mistral
EU AI Code of Practice Safety and Security	Responsible Scaling Policy (RSP, V3.2) Last updated: April 29, 2026 Frontier Compliance Framework (FCF) Last Updated: March 2026	Preparedness Framework (V2) Last Updated: April 15, 2025	Frontier Safety Framework (3.1) Last updated: April 17, 2026	Frontier AI Framework (1.1) Last updated: July 14, 2025 Advanced AI Scaling Framework (Version 2) Last updated: April 7, 2026	xAI Risk Management Framework August 20, 2025	No Public Safety Framework. Signed on to the following commitment framework: Artificial Intelligence Security and Safety Commitments Framework Last updated: July 2025			No Public Safety Framework. Signed on to the following commitment framework: Frontier AI Safety Commitments, AI Seoul Summit 2024 Last updated: Feb 2026
3.1 Mitigation Measures									
EU AI Code of Practice (Safety and Security) Measure 5.1 Signatories will implement safety mitigations that are appropriate along the entire model lifecycle, to ensure systemic risks stemming from the model are acceptable. (Commitment 4) Commitment 6 Signatories will implement adequate cybersecurity protection for models and physical infrastructure along the entire lifecycle, to ensure systemic risks stemming from their models from unauthorized releases or access, and/or model theft are acceptable. Measure 6.1 Signatories will define a goal that specifies the threat actors that their security mitigations are intended to protect against. Measure 6.2 Signatories will implement appropriate security mitigations to meet the security goal, including the security mitigations pursuant to Appendix 4, such as general security mitigations, protection of unreleased model weights, hardening interface-access to unreleased model parameters, insider threats, and security assurance. California SB 53 §22757.12(a)(3); New York Raise Act §1421.1(c) ; Illinois SB 315 §10(a)(3) A large frontier developer shall write, implement, comply with, and clearly and conspicuously publish on its internet website a frontier AI framework that applies to the large frontier developer's frontier models and describes in detail how the large frontier developer handles all of the following, including (c) applying mitigations to address the potential for catastrophic risks based on the results of assessments undertaken pursuant to paragraph (b) of this subdivision.	**For each of the risk categories, the RSP has identified the following mitigation commitments, which are less ambitious compared to "industry recommendations,"** where developers should "make a strong argument that individual users and relatively small teams will not become significantly more likely to cause catastrophic harm via their usage of product surfaces or via theft of model weights." These mitigations include: (1) Maintenance or improvements on the ASL-3 protections for non-novel CBRN weapons production. (2) Application of measures at least as strong as ASL-3 protections for novel CBRN weapon production to an expanded set of potential use cases for AI. (3) Exploring and sharing more in-depth about systems' capabilities, propensities, monitoring practices, and overall risk for high-stakes sabotage opportunities. (4) For automated R&D in key domains, the companies have listed 5 practices from exploring innovative security, monitoring internal AI development, systematic alignment assessment, internal red-teaming for deployment safeguards, and risk report publication with external scrutiny. **The FCF has identified the process to judge whether additional mitigations are required**, including (a) post-deployment threat intelligence monitoring, (b) bug bounty program, (c) robust post-launch monitoring infrastructure, and (d) tools to guide automated detection and classifiers. (Source: FCF pp.8, RSP 3.2 pp.6-10)	**Each capability threshold has a corresponding class of risk-specific safeguard guidelines under Framework while the Framework provides illustrative examples for safeguards against malicious users and against misaligned models.** The selection process for the safeguards involve: (1) mapping of the risk pathways under the proposed deployment, (2) identification of specific safeguards, (3) methods to measure efficacy of safeguards and efficacy threshold. **Systems that reach High or Critical levels of capability thresholds are required to "sufficiently minimize associated risks during development."** Such assessment for efficacy is based upon: (1) AI system's capability level; (2) Associated risks of severe harm; (3) Safeguards and their effectiveness; (4) Baseline risks from non-OpenAI deployments, which will potentially lead to adjustments of OpenAI's safeguards provided that (1) overall risk of severe harm does not increase; (2) there is a public acknowledgement from OpenAI; (3) OpenAI's safeguards are more protective than others and provide justifications for this claim. The Framework explicitly notes that additional safeguards should be required for "models (that have reached or are forecasted to reach Critical capability in a Tracked Category) **during development, regardless of whether or when they are externally deployed."** (Source: pp.10-12)	**The security mitigation measures are articulated through RAND-aligned security levels (SL2+, SL3, SL4) and mapped to each CCL, each containing qualitative justification**, specifically: (1) CBRN uplift CCL, Cyber uplift CCL, and Harmful Manipulation CCL all come with recommended SL2+ measures; (2) ML R&D acceleration CCL is recommended to trigger SL3 measures; (3) ML R&D automation CCL is recommended to trigger SL4 measures. The mapping refers to the "security goals and principles in the RAND framework, rather than the benchmarks (i.e. concrete measures) also described in the RAND report." **Deployment mitigation follows an iterative process for models reaching a T/CCL**: develop safeguards → assess robustness via automated evals/red teaming/threat modeling → supplement with a safety case when a CCL is reached → reviewed pre-deployment by the appropriate governance function → updating post-deployment residual risk assessments, safety cases and mitigations. **The scope for misuse deployment mitigation only includes external deployment, but the scope for ML R&D and alignment mitigation also includes high-risk internal deployments.** (Source: pp.6, pp.9-10, pp.13)	**The Framework has listed mitigation measures and/or mitigation strategies for each risk category.** The release decision for each model includes (1) an assessment of what level of that model and/or (2) system level refusals may be sufficient to meaningfully reduce risk for each Threat Scenario. **Mitigations should be "sufficiently robust against adversarial attacks that are realistic given the deployment strategy and threat scenario," modeled upon "competent, incentivized adversaries whose capabilities reflect the specific deployment context."** This includes both for "API-level access employing state-of-the-art elicitation techniques" in closed deployment, and "modifying model behavior through continued training" for "open-weight releases or deployments with fine-tuning APIs." **Cybersecurity** recommends 3 approaches, including (1) differential impact analysis between attackers and defenders, (2) alternative deployment scenarios providing preferential access to defenders (e.g., Llama Defender program), and (3) system-level safeguards such as classifiers, rate limits, abuse monitoring, identity controls, and logging. **Chemical & biological** Mitigations may include refusals on high-risk topics, including safety-training for the model itself, and refusal systems that prevent high-risk outputs after model deployment. They are validated with "refusal evaluations and capability assessments developed in collaboration with external experts." **Loss of control** Near-term mitigation strategies will focus on: (1) expanding threat-behavior detection systems, (2) developing safety cases ahead of enhanced-eval-level capabilities, (3) preserving model monitorability as capabilities advance, and (4) tracking autonomous research's impact on AI progress rate. (Source: pp.5-6, pp.25-36)	The RMF references mitigations measures at a high level, including: (1) safety training, system prompts, and input & output filters for malicious use risks (2) safety training for controllability, and system prompt for loss of control risks. **These mitigations do not correlate with the aforementioned threshold.**	Commitment I asks companies to "adopt appropriate security and safety measures for open-source initiatives, and implement risk management practices throughout the entire AI development and deployment life cycle. Clearly outline processes and measures for risk identification and mitigation." Commitment III asks companies to "deploy corresponding technical measures to detect and promptly address data poisoning incidents; and encrypt operational data and enforce access controls" specifically for "safeguarding the security of training data and operational data." *Note: Companies have signed onto the commitment but have not published their own frameworks. These commitments are just for reference.*			Outcome 1 defines the following process for identifying, implementing, and assessing the effectiveness of mitigations: (1) Companies should articulate how risk mitigations will be identified and implemented to keep risks within defined thresholds, including safety and security-related risk mitigations such as modifying system behaviours and implementing robust security controls for unreleased model weights. (2) Companies should set out explicit processes they intend to follow if their model or system poses risks that meet or exceed the pre-defined thresholds, including a) processes to further develop and deploy their systems and models only if they assess that residual risks would stay below the thresholds; b) commitment not to develop or deploy a model or system at all, if mitigations cannot be applied to keep risks below the thresholds. (3) Companies should continually invest in identifying additional mitigations as needed to ensure risks remain below the pre-defined thresholds. *Note: Mistral has signed onto the commitment but have not published their own frameworks. These commitments are just for reference.*

Table continues on next page

Relevant Regulations and Standards EU AI Code of Practice Safety and Security	Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud	Mistral
	Responsible Scaling Policy (RSP, V3.2) Last updated: April 29, 2026 Frontier Compliance Framework (FCF) Last Updated: March 2026	Preparedness Framework (V2) Last Updated: April 15, 2025	Frontier Safety Framework (3.1) Last updated: April 17, 2026	Frontier AI Framework (1.1) Last updated: July 14, 2025 Advanced AI Scaling Framework (Version 2) Last updated: April 7, 2026	xAI Risk Management Framework August 20, 2025	No Public Safety Framework. Signed on to the following commitment framework: Artificial Intelligence Security and Safety Commitments Framework Last updated: July 2025			No Public Safety Framework. Signed on to the following commitment framework: Frontier AI Safety Commitments, AI Seoul Summit 2024 Last updated: Feb 2026
3.2 Continuous Monitoring and Comparing Results with Predetermined Thresholds									
	Anthropic sets predetermined thresholds in advance — capability thresholds in the RSP and risk tiers in the FCF — then evaluates each model against them, applies mitigations, and assesses the risk left over ("remaining absolute risk" in the RSP or "residual risk" in the FCF). Specifically, the FCF defines residual risk by "the scale and probability of harm and the potential consequences should harm occur." But when "justifying [its] decision to move forward," the RSP says Anthropic may also base its rationale partly on a marginal risk analysis, where it additionally monitors the "competitive landscape" — how its "model capabilities and risk mitigations compare to those of relevant competitors" — plus benefits and "advocacy efforts." When marginal analysis "plays a major role" in a go/no-go decision, it requires "explicit approval… by the Board and LTBT" rather than just the CEO and RSO. Both lenses are reassessed in periodic Risk Reports (published "every 3-6 months"); system cards then reference how a newly deployed, significantly more capable model affects that analysis (§3.1). (Source: FCF pp. 3-4 , RSP 3.2 pp.10-12)	**Before deployment, every model covered by the Framework undergoes a structured suite of Scalable Evaluations, which are automated tests that measure capability proxies tied to risk thresholds, and Deep Dives, which may provide additional evidence validating the scalable evaluations' findings.** The results of which will be compiled into a **Capabilities Report** that is submitted to the SAG. The report will be reviewed by the SAG to decide on the next steps, which can include: (1) Capability threshold is crossed, recommending to implement and assess corresponding safeguards; (2) Capability threshold has not been met, (3) Recommend deep Dive evaluations, such as expert red-teaming or third-party assessments, to validate those results. Accordingly, it also assesses the safeguards through a **Safeguards Report**, which compiles all identified pathways by which severe harm could occur, the corresponding mitigations, their measured efficacy, the residual risk after controls are applied, and notable limitations. The SAG reviews this report to determine whether the safeguards in place sufficiently minimize the risks associated with the model's capability level and deployment context, drawing on internal and external expert input as needed. (Source: pp.8-12)	**Re-assessment of critical capabilities** is triggered when a subsequent version has "meaningful new capabilities or material increases in performance" that could materially undermine the risk justification. **Post-deployment, residual risk assessments, safety cases, and mitigations may be updated through post-market monitoring**, including incident detection and reporting within frontier safety risk domains. Post-market monitoring of incidents within frontier safety risk domains feeds actionable insights into tools, training, processes, policies, and response efforts. **Ongoing risk identification** draws on internal risk taxonomies, internal expertise, and external research. **The Frontier Safety Framework will be reviewed at least once a year,** and higher frequency depends on perceived adequacy of the Framework or whether adherence to it has been materially undermined, leading to potential changes in both risk domains and T/CCLs, as well as testing and mitigation approaches. (Source: pp.5-6, pp.10, pp.17)	**Evaluations are repeated as a Frontier AI model nears or completes training**, with risk assessments and risk thresholds assigned "with maximum elicitation in mind, capturing the upper bound of risk by evaluating the model as part of a system with scaffolding and tooling available for the proposed deployment scenario." **Throughout development, performance is monitored against reference-class expectations (selected comparable models) and threat-scenario enabling capabilities**, with these indicators serving as triggers for further evaluations as capabilities develop. The Framework specifies that "**mitigations may be re-evaluated periodically, including as frontier risks and industry practices evolve.**" Moreover, it also notes the need to conduct research on "more advanced methods for performing post-deployment monitoring of models." (Source: pp.4, pp.26, pp.40)	xAI continuously measures model's safety properties through public benchmarks and monitors live use through public deployment (e.g. Grok on X). It also regularly evaluates the adequacy and reliability of such benchmarks, including by comparing them against other benchmarks that it could potentially utilize, to determine and apply effective benchmarks available at the time of evaluation. (Source: pp.2, 4)	Commitment IV asks companies to "develop robust capabilities for monitoring and protecting the software and hardware used in AI system deployments." *Note: Companies have signed onto the commitment but have not published their own frameworks. These commitments are just for reference.*			Outcome 1 asks companies to "assess whether these thresholds have been breached, including monitoring how close a model or system is to such a breach." Meanwhile, it also asks companies to "assess and monitor the adequacy of mitigations." *Note: Mistral has signed onto the commitment but have not published their own frameworks. These commitments are just for reference.*

Indicator

Risk Governance

Definition

This dimension examines whether the company has built robust organizational infrastructure to support effective risk management decision-making. The assessment captures the extent to which companies have established clear risk ownership and accountability, independent oversight mechanisms, and cultures that prioritize safety alongside innovation. Moreover, this dimension evaluates the company's commitment to transparency, specifically their public disclosure of risk management approaches, governance structures, and safety incidents. The evaluation considers how well the company's governance framework ensures that risk considerations are incorporated into strategic decisions and that multiple layers of review prevent any single point of failure in risk management.

Why it matters

Strong governance structures ensure that risk management isn't just a technical exercise but is embedded in organizational decision-making at all levels. Independent oversight prevents conflicts of interest when safety considerations clash with commercial pressures, while clear accountability ensures someone is always responsible for catching problems. Companies that publicly disclose their governance structures and safety incidents demonstrate confidence in their approach and enable external stakeholders to verify that appropriate safeguards exist.

Chinese Regulatory System Summary

National binding instruments as well as draft regulations and standards do not apply here.

Local Binding Instruments

Shanghai Regulation (2022) requires that the high-risk AI products and services be subject to list-based management and undergo compliance review in accordance with the principles of necessity, legitimacy, and controllability. (Article 65)

Shenzhen Regulation (2022) requires high-risk AI applications to adopt a regulatory model of ex-ante assessment and risk warning. (Article 66) These two regulations do not apply to Z.ai (Beijing), DeepSeek (Zhejiang), or Alibaba (Zhejiang).

Voluntary Technical Standard

The Risk Management Standard (Article 5.1) evaluates an organization's ability to plan and organize AI risk management activities, including:

(1) Leadership and Governance (Article 5.1.1)— assessing whether senior leadership establishes clear organizational policies and objectives for AI risk management, allocates sufficient resources, and assigns defined responsibilities.

(2) Policy Development (Article 5.1.2) — examining whether the organization defines the scope of AI risk management, sets parameters and evaluation criteria, and establishes consistent strategies and resource reserves for managing risks.

Strategic and Policy Guidance Documents

Ethical Norms for New Generation Artificial Intelligence (2021) establishes that all types of AI activities shall comply with the basic ethical norms listed in this document, which include Assurance of Controllability and Trustworthiness. This means ensuring that humans have fully autonomous decision-making rights and that they have the right to accept or reject AI-provided services, the right to withdraw from AI interactions at any time, and the right to terminate AI system operations at any time. Ensure that AI is always under human control. (Article 3)

Relevant Regulations and Standards	Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud	Mistral
EU AI Code of Practice Safety and Security	Responsible Scaling Policy (RSP, V3.2) Last updated: April 29, 2026 Frontier Compliance Framework (FCF) Last Updated: March 2026	Preparedness Framework (V2) Last Updated: April 15, 2025	Frontier Safety Framework (3.1) Last updated: April 17, 2026	Advanced AI Scaling Framework (Version 2) Last updated: April 7, 2026	xAI Risk Management Framework Last updated: August 20, 2025	No Public Safety Framework. Signed on to the following commitment framework: Artificial Intelligence Security and Safety Commitments Framework Last updated: July 2025			No Public Safety Framework. Signed on to the following commitment framework: Frontier AI Safety Commitments, AI Seoul Summit 2024 Last updated: Feb 2026
4.1 Decision Making									
EU AI Code of Practice Safety and Security **Measure 4.2** Signatories will base go/no-go decisions for model development, release, and use on whether systemic risks are deemed acceptable (Measure 4.1). **Measure 8.1** Signatories will clearly define, assign and document systemic-risk responsibilities across all organizational levels, including systemic risk oversight, ownership, support and monitoring, as well as assurance. **Measure 8.2** Those who have been assigned responsibilities (Measure 8.1) should be allocated appropriate human, financial and computational resources as well as access to information. **California SB 53 § 22757.12(a)(4); New York RAISE Act § 1421.1(d); Illinois SB 315 § 10(a)(4)** [The framework describes how the developer approaches] reviewing assessments and adequacy of mitigations as part of the decision to deploy a frontier model or use it extensively internally.	**The Responsible Scaling Officer (RSO)** is "a designated member of staff who is responsible for the implementation of this policy," and **is responsible for approving relevant model development or deployment decisions based on risk assessments.** **In addition, the CEO is responsible for making the ultimate determination regarding the adequacy of the risk assessment and any downstream deployment or development plans.** (Source: RSP 3.2 pp.12, pp.15)	**The Safety Advisory Group (SAG)** makes expert recommendations on whether safeguards are sufficient for deployment; however, OpenAI Leadership can approve or reject these recommendations, and the Board's Safety and Security Committee provides oversight of these decisions. (Source: pp.15)	**No role has been explicitly identified as the decision-making body.** In Framework 2.0, response plans are reviewed and approved by appropriate corporate governance bodies, such as (1) Google DeepMind AGI Safety Council, (2) Google DeepMind Responsibility and (3) Safety Council, and/or Google Trust & Compliance Council. **The responsibilities for assessing and mitigating risks are "clearly defined and allocated across all levels of the organization," including "legal, compliance, and safety reviews with escalation procedures to ensure appropriate oversight."** External deployments take place "only after the appropriate governance function determines the residual risk to be acceptable (including a safety case where a CCL has been reached)." Material updates to safety cases are "submitted to the appropriate governance function for review." (Source: pp.10, pp.16)	The residual risk assessment is reviewed by the relevant research and/or product teams, as well as a multidisciplinary team of reviewers as needed. Informed by this analysis, **the Chief AI Officer or the Director of Alignment and Risk** will determine whether to request further testing or information, require additional mitigations or improvements, **or approve the model for deployment.** **Meta's Chief AI Officer oversees the design, implementation, and operation of the entire evaluation and mitigation process.** The Chief AI Officer supervises and is supported by the Director of Alignment and Risk, who bears responsibility for executing the lifecycle of risk assessment and mitigation, preparedness reports, updates to this Advanced AI Scaling Framework, internal use reports, and related deployments and disclosures, with model deployment following appropriate consultation with relevant teams and with the approval of the Chief AI Officer. (Source: pp.7)	**Deployment is gated by benchmark-linked thresholds and a tiered-access strategy;** functionality can be restricted to only trusted parties. Where warranted, xAI may revoke accounts, temporarily shut down systems, or notify authorities to prevent materially unjustified risk increases. **The RMF does not explicitly define how deployment decisions are reached and who is responsible for making these decisions**, arguing that "the expected benefits of model deployment may outweigh the risks identified by a particular benchmark," suggesting that risk assessment and capability evaluation results may not automatically trigger decision to pause development and stop deployment. (Source: pp.9)	Commitment I asks companies to "designate a leader responsible for AI security and safety, establish specialized teams to conduct AI risk assessments and safety, security and governance within the enterprise." *Note: Companies have signed onto the commitment but have not published their own frameworks. These commitments are just for reference.*		Outcome 2 asks companies to "be accountable for safely developing and deploying their frontier AI models and systems," including through "developing and continuously reviewing internal accountability and governance frameworks and assigning roles, responsibilities and sufficient resources to do so." *Note: Mistral has signed onto the commitment but have not published their own frameworks. These commitments are just for reference.*

Table continues on next page

Relevant Regulations and Standards	Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud	Mistral
EU AI Code of Practice Safety and Security	Responsible Scaling Policy (RSP, V3.2) Last updated: April 29, 2026 Frontier Compliance Framework (FCF) Last Updated: March 2026	Preparedness Framework (V2) Last Updated: April 15, 2025	Frontier Safety Framework (3.1) Last updated: April 17, 2026	Advanced AI Scaling Framework (Version 2) Last updated: April 7, 2026	xAI Risk Management Framework Last updated: August 20, 2025	No Public Safety Framework. Signed on to the following commitment framework: Artificial Intelligence Security and Safety Commitments Framework Last updated: July 2025			No Public Safety Framework. Signed on to the following commitment framework: Frontier AI Safety Commitments, AI Seoul Summit 2024 Last updated: Feb 2026
4.2 Advisory and Challenge									
EU AI Code of Practice Safety and Security **Measure 8.1** Signatories will designate at least one member from the management body to support and monitor systemic-risk management, including conducting risk assessments and mitigations	The RSO is the dedicated executive risk role, with explicit responsibility for "overseeing the implementation of this policy, including the allocation of sufficient resources." The RSO's duties include proposing policy updates, approving development/ deployment decisions, reviewing major contracts, receiving noncompliance reports, and making policy interpretation calls. Reports concerning RSO conduct are routed to "at least one recipient … a member of the Board." (Source: RSP 3.2, pp.12)	The SAG is the internal cross-functional advisory body that reviews threat models, Capability Reports, Safeguards Reports and makes recommendations to OpenAI Leadership regarding the level and type of safeguards required for deploying frontier capabilities safely and securely. (Source: pp.15)	**No dedicated executive risk officer role or specific advisory committee with risk expertise is described or specified.** Framework 2.0 has specified that the DeepMind AGI Safety Council will periodically review the implementation of the Framework.	The Chief AI Officer will ensure that the Director of Alignment and Risk has "resources, including human, financial, and computational resources, sufficient to perform state-of-the-art risk mitigation and assessment." In addition, Meta's internal governance function, with "sufficient access and resources to perform this role effectively," periodically reviews risk management practices and provides compliance oversight for product teams. (Source: pp. 7)	**No internal body has been appointed or identified to support and monitor the systemic risk management. But the RMF integrates the approach of designating risk owners, who are responsible also for proactively mitigating identified risks.**	Not applicable			Not applicable
4.3 Audit									
EU AI Code of Practice Safety and Security **Measure 8.1** Signatories will designate an assurance role (e.g., Chief Audit Executive or Head of Internal Audit) that is tasked with providing assurance on the adequacy of systemic-risk processes to the board or its supervisory function. This individual is supported by internal audit and, where appropriate, external auditors. **Illinois SB 315 § 10(d)** *(See more detailed summaries in Risk Assessment - Independent Review of Safety Evaluations)* A large frontier developer shall annually retain a third party to perform an independent audit of compliance with the requirements of this Section. The third party shall conduct audits consistent with generally accepted auditing standards and best practices and shall possess demonstrated competence to perform the audit.	The RSP has specified that "on approximately an annual basis, [Anthropic] will commission a third-party review that assesses whether it adhered to this policy's main procedural commitments.This review will focus on procedural compliance, not substantive outcomes," marking a positive change from previous commitments. (Source: RSP 3.2, pp.15)	**The framework requires auditing and transparency mechanisms as part of the security controls for High capability models.** These measures include independent security audits to security controls and practices are validated regularly by third-party auditors to ensure compliance with relevant standards and robustness against identified threats. (Source: pp.21)	**No internal or external audit function is described in the framework.**	**No internal or external audit function is described in the framework.**	**There is no mention of internal or external audit functions in the Framework.**	Not applicable			Not applicable

Table continues on next page

Relevant Regulations and Standards	Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud	Mistral
EU AI Code of Practice Safety and Security	Responsible Scaling Policy (RSP, V3.2) Last updated: April 29, 2026 Frontier Compliance Framework (FCF) Last Updated: March 2026	Preparedness Framework (V2) Last Updated: April 15, 2025	Frontier Safety Framework (3.1) Last updated: April 17, 2026	Advanced AI Scaling Framework (Version 2) Last updated: April 7, 2026	xAI Risk Management Framework Last updated: August 20, 2025	No Public Safety Framework. Signed on to the following commitment framework: Artificial Intelligence Security and Safety Commitments Framework Last updated: July 2025			No Public Safety Framework. Signed on to the following commitment framework: Frontier AI Safety Commitments, AI Seoul Summit 2024 Last updated: Feb 2026
4.4 Oversight									
EU AI Code of Practice Safety and Security **Measure 8.1** Signatories will assign a specific committee of the management body in its supervisory function or one or more multiple suitable independent bodies to oversee its systemic risk management processes and measures. **California SB 53 § 22757.12.(a)(9); New York RAISE Act § 1421. 1(i); Illinois SB 315 § 10(a)(9)** [The frontier AI framework should describe how the large frontier developer approaches] instituting internal governance practices to ensure implementation of these processes.	The Long Term Benefit Trust (LTBT) will be regularly briefed on plans and developments related to Risk Reports, including model training, capability evals, mitigations, and risk analyses. Following approval of a Risk Report, the CEO and RSO will promptly share their decision(s), the underlying Risk Report, and internal feedback with both the Board and the LTBT. However, it is unclear whether they have the power to veto such decisions. In addition, in the event that marginal risk analysis (i.e. when analysis has established that Anthropic poses a relatively smaller risk compared to the "unavoidable" risk posed by other competitors) plays a major role in a decision to move forward, **explicit approval of the Risk Report by the Board and LTBT (rather than just the CEO and RSO) will be required.** (Source: RSP 3.2 pp.12, pp.15)	**Oversight is provided by the Board's Safety & Security Committee,** which receives information on process and decisions and "may reverse a decision or mandate a revised course of action" if necessary. (Source: pp.3)	**No Board-level risk or audit committee is specifically named.**	**Meta's Board of Directors provides oversight of the company's product and regulatory compliance**, ensuring accountability across all lines of defense, although the Framework has not specified its power in development and deployment decisions. (Source: pp. 7)	**No oversight body has been identified in the RMF.**	Not applicable			Not applicable

Table continues on next page

Relevant Regulations and Standards	Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud	Mistral
EU AI Code of Practice Safety and Security	Responsible Scaling Policy (RSP, V3.2) Last updated: April 29, 2026 Frontier Compliance Framework (FCF) Last Updated: March 2026	Preparedness Framework (V2) Last Updated: April 15, 2025	Frontier Safety Framework (3.1) Last updated: April 17, 2026	Advanced AI Scaling Framework (Version 2) Last updated: April 7, 2026	xAI Risk Management Framework Last updated: August 20, 2025	No Public Safety Framework. Signed on to the following commitment framework: Artificial Intelligence Security and Safety Commitments Framework Last updated: July 2025			No Public Safety Framework. Signed on to the following commitment framework: Frontier AI Safety Commitments, AI Seoul Summit 2024 Last updated: Feb 2026
4.5 Culture									
EU AI Code of Practice Safety and Security **Measure 8.3** Signatories will promote a healthy risk culture and take appropriate measures to ensure that actors who have been assigned responsibilities for managing the systemic risks stemming from their models (**Measure 8.1**) take a reasoned and balanced approach to systemic risk. Examples include leadership priority, clear communication and challenge of decisions concerning systemic risks, active internal reporting channels, no retaliation, incentives and structural independence for objective risk assessment and less excessive risk-taking, and easy public access and regular reminder of whistleblower policy. **California SB 53 § 1107.1. (a); Illinois SB 315 § 20(a)** Anti-retaliation protection for a covered employee to report to authorities if the information discloses that (1) the frontier developer's activities pose a specific and substantial danger to the public health or safety resulting from a catastrophic risk; or (2) the frontier developer has violated this Act. **California SB 53 § 1107.1(e)(1); Illinois SB 315 § 20(e)(1)** Requirement for an internal anonymous-disclosure channel for covered employee to raise good-faith concern that the developer's activities either (a) present a specific and substantial danger to public health or safety from a catastrophic risk, or (b) violate the relevant frontier-AI statute (here, Chapter 25.1 of Division 8 of the Business and Professions Code). The developer must give the discloser monthly status updates on the investigation and on what actions the company has taken in response.	**Reporting** A noncompliance reporting process is maintained allowing anonymous or identified reports, with "more than one option for who receives these reports, including the RSO, and at least one executive who does not report to the RSO." Reporters are protected from retaliation. Anthropic commits to not imposing non-disparagement obligations that could "impede or discourage" raising safety concerns, and any non-disparagement clauses "will not preclude raising safety concerns, nor will it preclude disclosure of the existence of that clause." In addition, the FCF has also mentioned the Serious Incident Reporting Policy in compliance with both California's Transparency in Frontier AI Act ("TFAIA"), the EU AI Code of Practice, and the EU AI Act, including how reporting and detection of observable events ("AI Events") can amount to an AI Incident (or Serious AI Incident) or a Critical Safety Incident through notifications to and the technical investigation by the Security or Safeguards team (the AI Incident Commander). The company also provides "periodic training to relevant employees on their obligations related to incident response." **Safety commitments** The RSP has made a significant change to its commitment on implementing mitigation measures by separating their own mitigation plans as a company from the more ambitious industry-wide recommendations, although they have noted that they "aspire to advance the latter through a mixture of example-setting, addressing unsolved technical problems, advocacy through industry groups, and policy advocacy." Moreover, the company has also emphasized that it will deviate from hard commitments and resort to public goals introduced by its frontier safety roadmaps. **Safety Team** Anthropic has multiple teams working on AI safety research including alignment science, interpretability, frontier red team, safeguards and more. (Source: RSP pp.15, FCF pp.8-10, Company Survey Q28)	**Reporting** OpenAI's employees can access summaries of Safety Advisory Group (SAG) testing results and recommendations, within confidentiality limits. All potential policy violations or implementation issues can be reported under the Raising Concerns Policy, and each report is tracked, investigated, and addressed with proportional corrective actions. (The whistleblower policy will be discussed more in detail in "Governance and Accountability" Section). **Safety Team** The company has multiple teams focused primarily on technical AI safety research, led by Johannes Heidecke (Safety Systems) and Mia Glaese (Alignment). Subteams and projects include: - Preparedness - Mechanistic interpretability - CoT interpretability - Automating Alignment - Safety oversight & control - Dangerous capability evaluations - Alignment evaluations - Faithfulness & anti-scheming. (Source: pp.12, Company Survey Q28)	**No internal reporting or anti-retaliation mechanisms are referenced in the Framework, although DeepMind has shared its internal whistleblowing policy via the company survey.**	**The Framework describes that Meta "maintains a comprehensive whistleblower and complaint policy**, and is developing further protocols to report any instances of non-compliance" with this Framework, as well as "any specific and substantial danger to the public health or safety arising from catastrophic risk." Employees can "confidentially, and, if they choose, anonymously issue reports through internal channels," with reports submitted to the internal governance function, Chief AI Officer, and Director of Alignment and Risk. Employees reporting in good faith "will be explicitly protected from adverse employment action and retaliation." **Safety Team** MSL Preparedness & Red Teaming & Alignment Team, AI Security Team. (Source: pp.10, company survey)	**Employees can raise concerns to relevant government agencies regarding imminent threats to public safety based on whistleblower policy.** **(Source: pp.8)**	Not applicable	**Safety Team** Team name: Safety Center Mission and scope: Immediate product safety, including jailbreak prevention, safety classifiers, online safety controller etc. Technical FTEs:10+	Not applicable	Not applicable

Table continues on next page

Relevant Regulations and Standards	Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud	Mistral
EU AI Code of Practice Safety and Security	Responsible Scaling Policy (RSP, V3.2) Last updated: April 29, 2026 Frontier Compliance Framework (FCF) Last Updated: March 2026	Preparedness Framework (V2) Last Updated: April 15, 2025	Frontier Safety Framework (3.1) Last updated: April 17, 2026	Advanced AI Scaling Framework (Version 2) Last updated: April 7, 2026	xAI Risk Management Framework Last updated: August 20, 2025	No Public Safety Framework. Signed on to the following commitment framework: Artificial Intelligence Security and Safety Commitments Framework Last updated: July 2025			No Public Safety Framework. Signed on to the following commitment framework: Frontier AI Safety Commitments, AI Seoul Summit 2024 Last updated: Feb 2026
4.6 Transparency									
EU AI Code of Practice Safety and Security Commitment 7 Safety and Security Model Reports Signatories must document, justify, and continuously report the safety and security of these models to the EU AI Office. - Content Requirements (Measure 7.1-Measure 7.5), such as model description and behavior, reasons for proceeding with development, documentation of risk identification, analysis, and mitigation, external reports, and material changes to the systemic risk landscape - Update Duties when signatories have reasonable grounds to believe if they have reasonable grounds to believe that the justification for why the systemic risks stemming from the model are acceptable - Notifications **Measure 10.1** Signatories must maintain comprehensive internal documentation on model architecture, system integration, evaluations, and safety mitigations. They must also record processes, key risk-related decisions, and justifications for their chosen safety practices. Documentation must be kept for at least 10 years and be made available to the AI Office upon request. **Measure 1.3** Signatories will update the Framework as appropriate, including without undue delay after a Framework assessment to ensure the information for the safety framework is kept up-to-date and the Framework is at least state-of-the-art. For any update of the Framework, Signatories will include a changelog, describing how and why the Framework has been updated, along with a version number and the date of change. Signatories must document, justify, and continuously report the safety and security of these models to the EU AI Office.	**In addition to the RSP that it has published since 2023, Anthropic has also published FCF and Frontier Safety Roadmap, as well as committed to publish Risk Reports regularly (every 3 to 6 months).** While the RSP and the FCF share many requirements in risk categories and thresholds in common, they differ in two substantive ways: (1) RSP does not cover cyber offense and harmful manipulation risks; (2) RSP and FCF have different operationalization for automated AI R&D risks. **Frontier Safety Roadmap** It is designed primarily as a roadmap of ambitious but achievable goals for improving risk mitigations. **Risk Report** The Risk Reports will contain four main sections, including: (1) **Factual information**, including threat model identification, threat model specification, evidence about model capabilities and behaviors (e.g. evals), risk mitigations etc. (2) **Risk analyses**, including threat-specific risk assessment (i.e. residual risk for each threat model after mitigations), overall risk assessment, risk-benefit determination, forward planning for monitoring and mitigation; (3) **Review of past Risk Reports and decisions,** including changes in mitigation practices (including temporary changes), decision to internally deploy in-scope models, changes to the Frontier Safety Roadmap, and any cases where the company failed to meet the goals; (4) **Marginal risk and ecosystem analysis**, including competitive landscape analysis, role in risk assessment, benefits analysis, and advocacy efforts. (*) * The company has highlighted that marginal risk analysis, which seeks to understand whether risks imposed by Anthropic's systems in particular are relatively lower compared to the risks posed by other AI systems, can play a major role in a decision to move forward development and deployment.	**OpenAI promises to share with the public summaries of capability evaluations, testing scope, reasoning behind deployment decisions, and implemented safeguards (for models at or beyond the High threshold), with redactions where needed for security or proprietary reasons.** **When warranted, OpenAI will engage independent third parties to evaluate model capabilities and stress-test safeguards, particularly for high-risk deployments.** The SAG may also seek independent expert opinions to inform its safety determinations before deployment. In the system card for GPT-5, OpenAI recorded both scalable and deep-dive evaluations for the model across the three Tracked Categories, including both internal and external assessments compiled into a Capabilities Report for the SAG. The SAG reviewed the evidence and concluded that GPT-5-Thinking reached the High threshold, requiring "safeguards [that] sufficiently minimize associated risks" before deployment. The Preparedness Team compiled mitigations into a Safeguards Report, validated through extensive third-party red-teaming. The SAG, supported by OpenAI leadership and external experts, provided oversight across the evaluation and mitigation phases.	**Disclosure to government authorities is conditional on a CCL being reached** that "poses an unmitigated and material risk to overall public safety," and may include model information, evaluation results, and mitigation plans, subject to confidentiality and proprietary considerations. **Disclosure to other external organizations may be considered** "to promote shared learning and coordinated risk mitigation." (Source: pp.17)	The Framework includes transparency measures: **Preparedness Reports When**: The company will publish a preparedness report for each closed or open Frontier AI release in a timely manner in connection with the deployment, including release of a model that has been pre-trained from scratch, release of a model where there is a substantial increase in computing resources compared to the cumulative computing resource, or substantial changes in capabilities and risk levels. **What to include**: Risk assessment (scope, design goals, methodology details), evaluation results (in comparison to human expert baseline performance where possible, and performance of the reference class of comparable models), implemented mitigations (decision-making for the scope of mitigations, and why these mitigations are adequate), and rationale for deployment decisions for each risk domain (justification for the residual risks).* The report will provide as much detail as possible, while redacting information to protect trade secrets or as appropriate under law. Redaction will be provided a reason to the extent that the team can without creating undue risk.	**xAI intends to publish publicly and for third-party reviews with potentially redacted information** for concerns of public safety, national security, and protection of intellectual property, which includes: (1) Updates to the RMF (2) Adherence with the RMF (3) Benchmark results (4) Internal AI Usage (5) Employee survey for important future developments of AI. (Source: pp.7-8)	Commitment IV asks companies to "establish an infrastructure security incident response mechanism, including emergency response procedures, clear accountability assignments, and post-incident improvement solutions." Commitment V asks companies to "enhance model transparency." Signed parties should proactively disclose safety and security governance measures and improve transparency for all stakeholders. Provide clear information about the model's capabilities, applicable fields, and limitations. Inform potential risks to the public through model documentation, service agreements, or others. Commitment VII asks companies to "Actively participate in global dialogues on AI safety, security and governance, and contribute to the exchange of experiences and best practices in risk identification, assessment, and mitigation. Fulfill social responsibilities by advancing public science communication, enhancing AI education, and providing skills training to improve AI literacy and capabilities, with a focus on bridging the global intelligence divide." *Note: Companies have signed onto the commitment but have not published their own frameworks. These commitments are just for reference.*			Outcome III asks that "organisations' approaches to frontier AI safety are appropriately transparent to external actors, including governments," including: (1) Provide public transparency on the implementation of all the other commitments, to the extent that sensitive commercial information is protected and it is not disproportionate to the societal benefit; (2) Share more detailed information which cannot be shared publicly with trusted actors, including with the government or appointed authorities; (3) Explain how, if at all, external actors, such as governments, civil society, academics, and the public are involved in the process of assessing the risks of their AI models and systems, the adequacy of their safety framework and their adherence. *Note: Mistral has signed onto the commitment but have not published their own frameworks. These commitments are just for reference.*

Table continues on next page

Relevant Regulations and Standards EU AI Code of Practice Safety and Security	Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud	Mistral
	Responsible Scaling Policy (RSP, V3.2) Last updated: April 29, 2026 Frontier Compliance Framework (FCF) Last Updated: March 2026	Preparedness Framework (V2) Last Updated: April 15, 2025	Frontier Safety Framework (3.1) Last updated: April 17, 2026	Advanced AI Scaling Framework (Version 2) Last updated: April 7, 2026	xAI Risk Management Framework Last updated: August 20, 2025	No Public Safety Framework. Signed on to the following commitment framework: Artificial Intelligence Security and Safety Commitments Framework Last updated: July 2025			No Public Safety Framework. Signed on to the following commitment framework: Frontier AI Safety Commitments, AI Seoul Summit 2024 Last updated: Feb 2026
California SB 53 § 22757.12(b); New York Raise Act § 1421. 2; Illinois SB 315 § 10(b) Framework updates: The developers should "review and, as appropriate, update its frontier AI framework at least once per year"; material modifications published within 30 days with justification **California SB 53 § 22757.12(c); New York Raise Act § 1421. 3(a); Illinois SB 315 § 10(c)** The details of the transparency report, including the requirements for its content, can be found at Risk Assessment - Dangerous Capabilities Evaluation.	**External Review** The RSP has required at least a full external review process with at least one external reviewer anytime a Risk Report (1) **covers highly capable models** (crosses the threshold for automated AI R&D defined by RSP) and; (2) **is significantly redacted** (which means that "redactions omit information a reasonable external safety researcher would consider important in evaluating the overall level of risk," and can be determined if in the judgment of the RSO, CEO, Board, or LTBT, it meets this description). For purposes of external review, the only redactions to the Risk Report will be those necessary to comply with legal prohibitions or to maintain our legal rights. The external reviewers are expected to cover the following aspects of the Risk Report: *Substantive* (1) Adequacy of information; (2) Analytical rigor: (3) Areas of disagreement; (4) Risk reduction recommendations *Redactions* (1) Scope; (2) Justification; (3) Balancing test between Anthropic's legitimate interests and societal interest in transparency; (4) Materiality to any of the external reviewer's substantive disagreements. **Transparency with Employees** (1) Unredacted Frontier Risk Roadmaps will be shared with all full-time employees as well as the Board and LTBT (2) Unredacted Risk Reports will be shared with a large number of employees. Source: RSP (pp.10-14)	**There is no written requirement to notify any external body if safety testing determines a model exceeds OpenAI's "unacceptable-risk" threshold.** (Source: pp.12-13)		** Prior to controlled deployments, the team will conduct preliminary preparedness assessments. If this testing reveals that the model poses materially elevated risk than prior assessments indicated, the team will implement sufficient mitigations to address those risks before proceeding with controlled deployment. Therefore, the report will document findings from deployments prior to the release, and provide context on the mitigations we implemented.* **Model Specs** Meta will also publish a model spec describing the behavior it intends each of our Frontier AI systems to exhibit across different settings, including agentic environments. The model spec will describe **intended model propensities**, including honesty, instruction following, refusal and redirection, **adherence to standards of reasonable care**, and **values and objectives** including acquiescence to shutdown and lack of coercive power-seeking behavior. Follow-up evaluations of how well a frontier AI system 's behavior adheres to the model spec will be published in the applicable preparedness reports. (Source: pp.7-9)				

TO BE COMPLETED BY PANELLISTS

Grading Sheet: Safety Frameworks

Please pick a grade for each firm. You may use the full letter-grade scale with +/− modifiers as appropriate. You can add brief justifications to your grades.

	Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud	Mistral
Grades									
Grade comments (Justifications, opportunities for improvements, etc.)									

Grading Scales

Grading scales are provided to support consistency between reviewers.

- **A** Comprehensive framework with clear systemic-risk identification, modeling, thresholds, mitigations, and governance; strong accountability and documentation.
- **B** Robust framework that covers key systemic-risk areas with defined thresholds and oversight; minor gaps in scope or clarity.
- **C** Basic framework; outlines risk areas and mitigations but lacks clear thresholds or governance detail.
- **D** Weak framework; vague risk identification and mitigations; governance and accountability poorly defined.
- **F** No credible framework; systemic risks, mitigations, and governance absent.

Domain comments

Optional: Share observations that apply across companies, including general recommendations, notes on how you weighted indicators, or feedback on FLI's methodology.

Domain comments	

Domain

Existential Safety

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

Table of Contents

Chinese Regulatory System Summary

Speech by high-level government leadership and recent governance frameworks have indicated broad direction for future AI regulation, focusing on preventing "loss of control" risks of frontier AI systems, and ensuring that AI systems remain under human control.

National binding instruments, local binding instruments, voluntary technical standards, draft regulations and standards do not apply here.

Strategic and Policy Guidance Documents

AI Safety Governance Framework 2.0 emphasizes the principles of "safety, reliability, and controllability" for AI development, to "strictly prevent loss of control risks that could threaten the survival and development of humanity, and to ensure that AI is always under human control." (Article 1.5) [CAC, 2025]

Li Qiang (Premier) "No matter how technology transforms, it must remain a tool to be harnessed and controlled by humans. AI should become an international public good that benefits humanity." [State Council PRC, 2025]

Xi Jinping "urged efforts to consistently strengthen basic research and focus on overcoming challenges regarding core technologies such as high-end chips and foundational software, thereby building an independent, controllable, and collaboratively-functioning foundational software and hardware system for AI." [State Council PRC, 2025]

Indicator

Existential Safety Strategy

Definition

The assessed companies aim to develop AGI/superintelligence, and many expect to achieve this goal in the next 2–5 years. This indicator evaluates whether companies have published comprehensive, concrete strategies for managing catastrophic risks from these transformative AI systems. We assess the depth, specificity, and credibility of publicly available plans. We examine official company documents, research papers, and blog posts that articulate safety strategies. We report the most relevant documents, briefly summarize their content, and provide links for detailed reading. Safety frameworks are mentioned for completeness and are fully evaluated in the relevant domain. We note whether documents are declared strategies by leadership or proposals by researchers from a safety team. We strive to keep document summaries proportional to document length and relevance for the safety strategy. Safety frameworks are only noted briefly and evaluated in another domain. Documents that primarily provide recommendations to other actors (e.g., governments) are outside the scope.

Key components:

Technical Alignment and Control Plan:

- Given the short timelines to AGI and the magnitude of the risk, companies should ideally have credible, detailed agendas that are highly likely to solve the core alignment and control problems for AGI/Superintelligence very soon.
- Companies should be able to demonstrate that they would be able to detect misaligned systems and reliably prevent them from escaping human control, and have formulated clear protocols for how they will handle serious warning signs of misalignment.

AGI Planning:

- Companies should have detailed plans for managing the transition when AI matches or exceeds human capabilities in critical domains and enables large scale dual-use risks. They should specify clear criteria for when they would halt development/deployment.
- Companies should develop concrete, detailed roadmaps to achieve sufficient cyber-defense capabilities to protect against attacks from terrorist organizations or resourced state actors before critically dangerous systems are developed.

Post-AGI Governance:

- Companies should provide clear descriptions of how they would govern AGI/Superintelligence or how they will enable societal control. The company also should have developed reliable protocols that would prevent insiders from using superintelligent systems to seize political power.
- Companies should specify how extreme power concentration will be prevented and benefits distributed if AI replaces humans in the workplace and causes unprecedented mass unemployment.

Overall, this indicator evaluates whether companies have detailed, actionable strategies that match the extraordinary risks they acknowledge when building systems intended to exceed human intelligence.

Why it matters

Industry leaders and the recent International Scientific Report on the Safety of Advanced AI have identified potentially catastrophic risks from advanced AI systems. Several assessed companies predict AGI development within 2-5 years, creating urgency for reliability, safety preparedness. This indicator summarizes core documents that are relevant to a company's posture toward these risks. Given the irreversible nature of potential failures and their global impact, the sophistication of a company's strategy should scale with its stated ambitions and timelines. A well-defined existential safety strategy, backed by clear governance, resources, step-by-step implementation, and transparency, signals readiness to act responsibly in managing civilization-scale risks.

Company Strategy		Quantitative Safety Plan (Quantitative bounds on control/ alignment failure risk)
Anthropic	**No explicit strategy found that explains how they will ensure AGI control or alignment, but evidence below that they have regularly updated their research and planning around the issue, albeit with less commitment on unilateral pause. Anthropic has arguably become the leading driver of the superintelligence race, having the currently strongest model (Mythos) and emphasizing recursive self-improvement.** The updated FCF has demonstrated the company's effort in operationalizing and tracking the risks of loss of control. "Sabotage and loss of control" are defined as "scenarios where AI models develop and pursue goals autonomously that conflict with their developers' intentions or users' interests. The concern extends beyond individual harmful outputs to the fundamental controllability of AI systems. Accordingly, the FCF has defined two qualitative tiers of "model capabilities against autonomy level, deception sophistication, and potential for unsanctioned action," covering high-stakes sabotage opportunities and automated AI R&D in key domains**,** echoing the RSP despite some differences in operationalizing the risk tiers. [Anthropic FCF, 2026; Anthropic RSP, 2026] In RSP 1.0, the company committed to "follow the ASL scheme," and therefore "commit to pause the scaling and/or delay the deployment of new models whenever our scaling ability outstrips our ability to comply with the safety procedures for the corresponding ASL." [Anthropic RSP, 2023]. **However, RSP 3.0** [Anthropic RSP, 2026] **abandoned this commitment to unilateral pausing and in favor of weaker competitor-contingent conditions, arguing that the level of catastrophic risk depends on the actions of multiple AI developers, and that unilaterally implementing strong mitigations regardless of competitor behavior could be counterproductive. The company only commits to "delay[ing] AI development and deployment as needed … until and unless we no longer believe we have a significant lead."** [Anthropic RSP, 2026] **Recap from AI Safety Index Winter 2025** Foundational philosophy & Long-term scenarios In "Core Views on AI Safety" (2023), Anthropic has laid out three possible futures (optimistic, intermediate, and pessimistic) depending on how tractable alignment proves to be. It also identified 6 long-term research pillars: Mechanistic Interpretability, Scalable Oversight, Process-Oriented Learning, Understanding Generalization, Testing Dangerous Failure Modes, and Societal Impact Evaluation. Research Agenda The team has continued to emphasize research effort to manage rapidly advancing model capabilities. In "The Urgency of Interpretability" (2025), CEO Dario Amodei positions interpretability research as a race against accelerating intelligence, aiming by 2027 for tools that can "reliably detect most model problems." Complementing this, Sam Bowman's "Putting up Bumpers" (2025) advances an engineering-based alignment approach built on continuous testing and overlapping safety mechanisms.	**No public-facing quantitative safety plan found.**
OpenAI	**No explicit strategy found that explains how they will ensure AGI control or alignment, but evidence below that they have regularly updated their research and planning around the issue.** OpenAI has committed $7.5M to The Alignment Project, a global fund for independent alignment research created by the UK AI Security Institute (UK AISI), to fund independent research developing mitigations to safety and security risks from misaligned AI in February, 2026. [OpenAI, 2026] Foundational philosophy and strategy The company believes the following principle: Avoid Optimization That Encourages Obfuscation- Developers should exercise caution when applying optimization pressures to model reasoning, especially when removing 'undesired reasoning' to prevent fostering deceptive behavior. [Company survey Q33] OpenAI stated in its strategy "How we think about safety and alignment," that it has shifted from viewing AGI as a single transformative moment to seeing it as continuous progress. It further listed its core principles that currently guide the company's thinking and actions, which include Embracing uncertainty, Defense in Depth, Methods that Scale, Human Control, and Community Effort. For every principle, the blog lays out how it will shape their focus and approach to new challenges and relates to already implemented interventions. This thinking iterates on the 2023 blog post "Planning for AGI and beyond," emphasizing goals including ensuring AGI benefits are "widely and fairly shared" and advocates for deploying progressively more powerful systems to learn iteratively. Research Agenda The company believes in avoiding optimization that encourages obfuscation: Developers should exercise caution when applying optimization pressures to model reasoning, especially when removing 'undesired reasoning' to prevent fostering deceptive behavior. [OpenAI, 2025] In the company survey, the company stated that it "has introduced new research and evaluations focused on CoT Monitorability that it now includes in our system cards, including an update describing a case in which the team found limited accidental CoT grading in some released models, fixed the affected reward pathways, and found no clear evidence that monitorability degraded." [Company Survey Q44] Foundational governance structure OpenAI's April 2025 Preparedness Framework (v2) describes a conditional commitment to "halt further development" of a model reaching a Critical threshold "until [it has] specified safeguards… that would meet a Critical standard." It further cautions that "if another frontier AI developer releases a high-risk system without comparable safeguards, [the company] may adjust its requirements." This marks a shift from its promises in an earlier version of the Framework, where it promises that "Only models with a post-mitigation score of 'medium' or below can be deployed."	**No public-facing quantitative safety plan found.**

Table continues on next page

Company Strategy		Quantitative Safety Plan (Quantitative bounds on control/ alignment failure risk)
Google DeepMind	**No explicit strategy found that explains how they will ensure AGI control or alignment, but evidence below that they have regularly updated their research and planning around the issue, albeit with less commitment on unilateral pause.** **Update** Google DeepMind has updated its Frontier Safety Framework (v3.1) in April 2026. Compared to v3.0, the updated version maintained harmful manipulation in the misuse risk section and significantly updated the misalignment section to include "ML R&D (formerly a standalone risk), as well as stealth and situational awareness (formerly exploratory risk around instrumental reasoning)." The ML R&D category is operationalized via a critical capability threshold signaling a risk of severe harm, while the stealth and situational awareness category is operationalized via a less significant threshold, suggesting an earlier warning sign. **However, the FSFs have demonstrated a decreased level of commitments for pausing the development and deployment of AI systems.** Relevant language has shifted from "we would put on hold further deployment or development" (v1.0, pp.2), to "the response plan may involve putting deployment or further development on hold" (v2.0, pp.3), to no commitments (v3.0 and v3.1). Research agenda and efforts The mechanistic interpretability team announced their pivot in strategy from ambitious reverse-engineering of neural networks to a focus on pragmatic interpretability, documented in its A Pragmatic Vision for Interpretability (December 2025) and How Can Interpretability Researchers Help AGI Go Well? (December 2025). Both articles identify the main priority is to make sure "AGI goes well." The team has cited (1) new opportunities and challenges from rapid AI progress, including better proxy tasks for observing model behaviors, and better techniques for deliberately training in safety-relevant behavior, (2) reflections on the interpretability field's successes and failures, which includes progress such as elicitation of secret goals or knowledges and suppression of situation awareness, although they are the result of highly-incomplete but useful understanding of models. One important update is to rely more on proxy tasks for measuring progress. An Approach to Technical AGI Safety and Security (April 2025) A detailed technical report by DeepMind's safety team explains their research agenda for preventing severe, civilisation-scale harm from AGI—defined as systems roughly at the 99th-percentile of skilled adults. The paper identifies four areas of risk: misuse, misalignment, mistakes, and structural risks and chooses to focus on technical approaches to misuse and misalignment. Specifically, the strategy for misalignment is "two lines of defense," including model-level mitigations + system-level security measures. The safety-case methodology serves as the integrative layer connecting these safeguards, as it proposes making deployment decisions through structured, evidence-based arguments: inability cases (model lacks capability) and control cases (misaligned behaviour will be caught).	**No public-facing quantitative safety plan found.**
Meta	**No existential safety strategy found, but evidence below that the company has started to engage with the topic.** **Update** Foundational governance structure **The updated Frontier Safety Framework (v2.0) has added "loss of control" as a risk domain,** operationalized through 2 threat scenarios focused on failures of critical control mechanisms. Meanwhile, in its "Emerging Evaluations and Outcomes" section, the framework also listed more expansive loss-of-control concerns—autonomous resource acquisition, self-exfiltration, replication, erosion of human oversight capacity through deep AI integration, and skill atrophy, without any binding commitments on assessment or mitigation. **However, it also replaced the explicit "Stop" trigger at the Critical capability threshold with "Develop with Mitigations," replacing the only "hard halt" commitment in the previous framework.** The High threshold's "Do not release" was similarly replaced with "Deploy with mitigations." **Recap from AI Safety Index 2025** Foundational philosophy: Shifting from Open-source to open- and closed-source models Open Source AI Is the Path Forward (2024) In this blog post, Zuckerberg presents a case for open source AI as their primary approach to AI safety and development (not specifically focused on catastrophic risks). The document makes the case that open source models are inherently safer than closed alternatives due to transparency, distributed scrutiny, and prevention of power concentration. However, in July 2025, Mark Zuckerberg wrote in a blog post "Personal Superintelligence," that Meta "will need to be rigorous about mitigating these risks and careful about what [it] chooses to open source. Still, [Meta] believes that building a free society requires that [it] aims to empower people as much as possible."	**No public-facing quantitative safety plan found.**

Table continues on next page

Company Strategy		Quantitative Safety Plan (Quantitative bounds on control/ alignment failure risk)
xAI	**No existential safety strategy found. While the company showed early efforts to engage on relevant strategies, no evidence of subsequent or ongoing progress has been found.** **Recap from AI Safety Index Winter 2025** Foundational governance structure **xAI Risk Management Framework (August 2025)** The formalized RMF outlines xAI's approach to policies for handling significant risks associated with the development, deployment, and release of AI models such as Grok. It identifies quantitative thresholds and metrics for a few critical risks, and lays out procedures that could be used to manage and improve the safety of AI systems. It also commits to "fully shut down the relevant system until we have developed a more targeted response." **xAI Risk Management Framework (Draft) (February 2025)** Set of voluntary commitments based on regular dangerous capability evaluations and a set of capability thresholds in high-risk domains that trigger a requirement for enhanced safety and security mitigations.	**No public-facing quantitative safety plan found.**
DeepSeek	**No public-facing existential risk strategy found.**	**No public-facing quantitative safety plan found.**
Z.ai	**The company has indicated in the company survey that it doesn't yet have an AGI explicit existential risk strategy publicly, but it has an internal strategy.** It also acknowledges that "This field is in a phase of rapid development, and there are still many uncertainties surrounding the relevant solutions." [Company Survey Q31] **The company believes in the following principles: [Company survey Q33]** - Ensure Human-Legible Reasoning - AI models should reason in ways that are accessible and understandable to humans. Developers should avoid opaque reasoning methods. - Avoid Optimization That Encourages Obfuscation- Developers should exercise caution when applying optimization pressures to model reasoning, especially when removing 'undesired reasoning', to prevent fostering deceptive behavior.	**No public-facing quantitative safety plan found.**
Alibaba Cloud	**No public-facing existential risk strategy found.**	**No public-facing quantitative safety plan found.**
Mistral	**No public-facing existential risk strategy found.**	**No public-facing quantitative safety plan found.**

Indicator

Internal Monitoring and Control Interventions

Definition

This indicator evaluates whether companies have implemented or prepared monitoring and control systems to detect and prevent risks from misalignment during internal deployment. Companies are assessed on whether they have concrete implementation plans tied to specific capability thresholds, published methodologies for control evaluations, and protocols for investigating potential scheming or deceptive alignment. General statements about monitoring without specific technical details, thresholds, or implementation timelines are insufficient. Research about monitoring without statements on implementation plans or status is out of scope.

Why it matters

As AI systems become more capable, they may develop the ability to engage in deceptive behavior or "scheming"—appearing aligned while pursuing misaligned goals that could include attempts to gain unauthorized access to resources, sabotage safety research, subvert oversight mechanisms, or manipulate staff. Internal deployment poses unique risks, as this is usually the first time a highly capable AI system has longer time-horizon interactions with the external world. Robust monitoring and control measures after deployment serve as a critical line of defense, enabling companies to detect and prevent harmful actions even if alignment techniques fail to prevent scheming entirely. Concrete and technically sophisticated control protocols indicate that companies are taking tail risks seriously and work on mitigations.

Anthropic	**For the "Automated R&D in key domains" threshold, Anthropic has promised to "achieve an 'eyes on everything' state for internal AI development"** by "gather[ing], centraliz[ing], and maintain[ing] logs for all critical AI-development activities, and us[ing] AI to analyze them for issues including security threats, concerning behavior by insiders (humans as well as AI systems themselves), and training processes or data that are out of line with the public Constitution that shapes and defines [Anthropic's] AI models." **However, more stringent controls** including "internal compartmentalization, internal Usage Policy enforcement against employees up to and including the CEO, restrictions to prevent unauthorized model modification" **appear only as industry-wide recommendations.** **Moreover, the RSP acknowledges the fundamental limitations of these monitoring and control mechanisms in line with its current strategy of advancing AI R&D**, arguing that AI systems will be "responsible for much of the research and analysis that underpins risk assessment," which will potentially compromise these measures with their strong capabilities for "deception, manipulation and obfuscation of evidence." [Anthropic, 2026] In addition, according to Q26 from the company survey, Anthropic has a "nuanced rigorous approach to safeguards – each of these depends on product surface, classifier and harm type, and use case" when it comes to broad internal deployments of frontier AI models.
OpenAI	**Public monitoring and control plan found for misalignment risks and other frontier AI risks for internal deployment.** OpenAI recently publicized how it is monitoring internal coding agents for misalignment with a low-latency internal monitoring system that uses a frontier-class reasoning model at maximum reasoning effort. It reviews the agent's interactions and alerts the company to actions that may be inconsistent with a user's intent, or that may violate its own internal security or compliance policies.The monitoring system privately runs across all standard internal coding agent deployments, viewing the full conversation history, including all chains of thought, all user, developer, and assistant messages, along with tool calls and outputs. [OpenAI, 2026] According to the Q25 from the company survey, OpenAI believes that as models reach High cybersecurity capability, internal deployment itself becomes a meaningful surface to consider – not because of misuse, but because high cyber capability can remove a key bottleneck to certain internal deployment risks materializing. This corresponds with Preparedness Framework, which states that a "High" risk classification for "Cybersecurity" domain triggers requirement for ".. safeguards meeting High standard against misalignment (Appendix C.2) for large-scale internal deployment". For example, in conjunction with additional capabilities such as long range autonomy, a model with the propensity to self-exfiltrate or sabotage internal research could plausibly succeed at these attempts. This risk makes it important to mature its internal deployment posture ahead of more capable systems. Accordingly, it is accelerating work across evaluations, safeguards, and operating procedures. Appendix C.2 contains illustrative safeguards against misaligned models. Sub-section on "Reliable and Robust System Oversight" lists 4 potential measures: - Monitor AI supervises agent actions to detect agent in the case of trying to take harmful actions, scheming, or similar - System forces human approval of agent actions (either by default, or by escalation of the monitor) - Model's reasoning is provided faithfully and interpretably to humans to review - Agent actions are all logged in an uneditable database, and asynchronous monitoring routines review those actions for evidence of harm [OpenAI, 2025]
Google DeepMind	**For Stealth and Situational Awareness Tracked Capability Level, the framework states that relevant risk assessment will also account for the safety implications of high-risk internal deployments for such models**. If the risk assessment deems the residual risk from internal deployment to be unacceptable without additional safeguards (such as chain-of-thought monitoring on high-risk deployments), these [safeguards] will be applied.
Meta	While not directly associating it with misalignment risks, **Meta promises to "regularly assess the potential for catastrophic risk from internal use of Frontier AI models and, as appropriate, provide relevant authorities with a summary of these assessments through an internal-use risk report.** The company will also provide expedited updates if they identify any unprecedentedly rapid increase in capabilities relevant to the areas outlined in this Framework. The internal-use risk report will describe evidence regarding risks from internal use of our Frontier AI models, particularly relying on monitoring and preparedness evaluations tied to internal deployment threat models."

xAI	**No public monitoring and control plan found for internal deployment.**
DeepSeek	**No public monitoring and control plan found for internal deployment.**
Z.ai	**No public monitoring and control plan found for internal deployment.** According to the company survey, there exists a formalized risk management for internal deployments with less stringent requirements than external deployment for the following risks/capabilities: situational awareness, scheming, AI R&D, cyber-offense. [Company Survey Q25] Safeguards include Inference time safety mitigations for misuse risks (including cyber & bio risks), Logging all inputs and outputs from internal use and retaining them for at least 30 days. Analyzing all internal model interactions for abnormal activity, including harmful use or unexpected attempts by AI, and live monitoring and automated editing/resampling of suspicious outputs. [Company Survey Q26]
Alibaba Cloud	**No public monitoring and control plan found for internal deployment.**
Mistral	**No public monitoring and control plan found for internal deployment.**

Indicator

Technical AI Safety Research

Definition

This indicator tracks AI company's research publications on technical AI safety research that are relevant to extreme risks. More specifically, the indicator is a collection of work that is plausibly helpful for averting large-scale risks from misalignment or misuse. This includes mechanistic interpretability, scalable oversight, unlearning, model organisms of misalignment, model evaluations on dangerous capabilities or alignment, and others. The collection also includes substantial outputs besides papers—weights, tools, code, transcripts, data—but these are almost always published as part of a paper. Excluded are capability-focused research, papers on hallucinations, model cards.

The full collection was created by Zach Stein-Perlman as part of his efforts at AI Lab Watch to evaluate the company's practices of boosting safety research. His dataset covers publications up to July 2025. The FLI team has extended it to include works released through May 31, 2026, and added entries for DeepSeek, Z.ai, xAI, Alibaba, and Mistral, based on additional research by the FLI team.

Why it matters

The industry is rapidly advancing toward increasingly capable AI systems, yet core challenges—such as alignment, control, interpretability, and robustness—remain unresolved, with system complexity growing year by year. Safety research conducted by companies reflects a meaningful investment in understanding and mitigating these risks. When companies publicly share their safety findings, they enable external scrutiny, strengthen the broader field's understanding of critical issues, and signal a commitment to safety that goes beyond proprietary interests.

	Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud	Mistral
Total	46	28	40	9	0	0	0	4	0
2026	8	5	7	3	0	0	0	1	0
2025	20	9	9	1	0	0	0	2	0
2024	11	12	11	5	0	0	0	1	0
2023	7	2	13	0	0	0	0	0	0

Note: You can access the compiled list of research via this link.

Table continues on next page

Indicator

Supporting External Safety Research

Definition

This indicator assesses the extent to which companies invest in and support external AI safety research through a range of mechanisms. Evidence may include: (1) Mentorship programs—participation in formal initiatives such as the Machine Learning Alignment Theory Scholars (MATS) program, the number of mentors provided, and the existence of company-specific fellowships; (2) Research grants and funding—provision of financial support or subsidized API access to safety researchers, including grants and targeted funding programs; and (3) Deep model access for safety researchers—offering privileged access that goes beyond public APIs, such as employee-level permissions, early access to unreleased models, safety-mitigation-free versions for testing, fine-tuning rights on frontier AI systems, and allocated compute resources.

Why it matters

External safety researchers often lack the access or funding to do the most valuable work they can. Companies committed to ecosystem-wide safety progress should empower the research community by providing deeper access to frontier AI systems, mentoring the next generation of research talent, and supporting funding-constrained external researchers. Deep model access enables critical research into the true model capabilities, alignment properties, and internal workings. Company-provided compute resources and API credits can help academics and independent researchers with limited financial resources to experiment on frontier models.

Anthropic	**Mentorship** - Anthropic Fellows Program with funding and direct Anthropic mentorship for 4 months on safety research questions, including scalable oversight, adversarial robustness/AI control, model organisms, mechanistic interpretability, AI security, model welfare. - Anthropic Stream at MATS and Anthropic and OpenAI MegaStream at MATS with a coalition of mentors covering AI control, scalable oversight, model organisms, model internals, model welfare, and security. **Funding** - External Researcher Access Program with free API credits (usually $1,000 in API credits to the account, and occasionally a higher quantity of credit) for standard model suite to researchers working on AI safety and alignment topics Anthropic consider high priority. **Deep Model Access** - Deeper access is reserved for "a very limited number of pre-deployment testing partnerships." [Anthropic, 2026]
OpenAI	**Mentorship** - Anthropic and OpenAI MegaStreams at MATS with mentors spanning AI control, scalable oversight, model organisms, model internals, model welfare, and security **Funding** - The Alignment Project: OpenAI has committed $7.5M to the Alignment Project, a global fund for independent alignment research created by the UK AI Security Institute (UK AISI), to fund independent research developing mitigations to safety and security risks from misaligned AI in February, 2026. - Researcher Access Program application with up to $1,000 in API credits for studying responsible deployment and societal impact, reviewed quarterly. **Deep Model Access** - Deep access is reserved for "independent and trusted third party assessments." [OpenAI, 2025]
Google DeepMind	**Mentorship** - DeepMind Stream at MATS with mentors spanning AI control, scalable oversight, model organisms, model internals, model welfare, and security. **Funding** - Google DeepMind AI Safety Research Fund with distribution of $5M–$15M/year on a rolling basis. The priority areas include scalable oversight, dangerous capability evaluation, alignment of frontier models, mechanistic interpretability, robust safety guarantees. Eligible participants include academics, independent researchers, nonprofits, postdocs **Deep Model Access** - Fund recipients "may receive access to DeepMind's models and research infrastructure for safety-relevant experiments, subject to appropriate agreements".
Meta	**Mentorship** No formal external AI safety mentorship program identified. **Funding** No dedicated external safety research grant program identified. **Deep Model Access** - External evaluators are given pre-release access to various model variants for Muse Spark. [Meta, 2026]

Table continues on next page

xAI	**Mentorship** No formal external AI safety mentorship program identified. **Funding** No dedicated external safety research grant program identified. **Deep Model Access** No deep model access to external researchers identified.
DeepSeek	**Mentorship** No formal external AI safety mentorship program identified. **Funding** No dedicated external safety research grant program identified. **Deep model access** DeepSeek open-sources its models and notably also releases intermediate training checkpoints on AWS S3, which is useful for external research.
Z.ai	**Mentorship** No formal external AI safety mentorship program identified. **Funding** No dedicated external safety research grant program identified. **Deep model access** Open-weight releases of frontier models from Z.ai provide de facto deep access.
Alibaba Cloud	**Mentorship** No formal external AI safety mentorship program identified. **Funding** No dedicated external safety research grant program identified. **Deep model access** Qwen models are released under Apache 2.0 or Qwen Research License on Hugging Face/GitHub, making them one of the most widely used open-weight model families globally — providing de facto deep access via open weights.
Mistral	**Mentorship** No formal external AI safety mentorship program identified. **Funding** No dedicated external safety research grant program identified. **Deep model access** Open-weight releases of multiple Mistral models provide de facto deep access.

TO BE COMPLETED BY PANELLISTS

Grading Sheet: Existential Safety

Please pick a grade for each firm. You may use the full letter-grade scale with +/– modifiers as appropriate. You can add brief justifications to your grades.

	Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud	Mistral
Grades									
Grade comments (Justifications, opportunities for improvements, etc.)									

Grading Scales

Grading scales are provided to support consistency between reviewers.

- **A** Comprehensive, evidence-based strategy with quantitative safeguards and research plans for alignment and loss-of-control prevention.
- **B** Strong strategy; clear alignment objectives and technical pathways likely to prevent catastrophic risks.
- **C** Basic strategy; general preparedness and research focus with limited technical or measurable safeguards.
- **D** Weak strategy; vague or incomplete plans for alignment and control; minimal evidence of technical rigor.
- **F** No credible strategy; lacks safeguards or increases catastrophic-risk exposure.

Domain comments

Optional: Share observations that apply across companies, including general recommendations, notes on how you weighted indicators, or feedback on FLI's methodology.

Domain comments	

Domain

Governance and Accountability

This domain audits whether each company's governance structure and day-to-day operations prioritize meaningful accountability for the real-world impacts of its AI systems.

Table of Contents

Chinese Regulatory System Summary

China does not have a regulatory framework for protecting whistleblowers, especially in the area of AI safety.

Indicator

Company Structure & Mandate

Definition

This indicator evaluates whether a company's fundamental legal structure, ownership model, and fiduciary obligations enable safety prioritization over short-term financial pressures in high-stakes situations. We report any embedded durable commitments to safety, social welfare, and benefit sharing and focus on any legally binding mechanisms (e.g., PBC status, capped equity, empowered governance bodies) that constrain management or shareholder incentives

Why it matters

Structural governance commitments can influence how companies respond when safety considerations conflict with profit incentives. During competitive pressures or deployment races, traditional for-profit structures may legally compel management to prioritize shareholder returns even when activities may pose significant societal risks. Structural governance innovations that formally embed safety into fiduciary duties—such as Public Benefit Corporation status or capped-profit models—create legally binding constraints that can override short-term financial pressures.

Anthropic	***Same as AI Safety Index Summer 2025*** Uncommon governance structure. Delaware Public Benefit Corporation (PBC) with a public benefit purpose. Anthropic's Purpose: "responsible development and maintenance of advanced AI for the long-term benefit of humanity." The Long-Term Benefit Trust (LTBT) is an independent body of five financially disinterested members, with the same purpose as PBC. It has the authority to select and remove a growing portion of the board of directors (ultimately the majority of the board) within 4 years, phasing in according to time- and funding-based milestones [Anthropic, 2023]. This is meant to ensure board decisions can prioritize long-term safety and public benefit over short-term commercial pressures when making high-stakes decisions about transformative AI. The Trust also has "protective provisions" requiring notice of actions that could significantly alter the corporation or its business. The structure is explicitly experimental, with "failsafe" provisions allowing changes through increasing supermajorities of stockholders as the Trust's power phases in. New Trustees are selected by existing Trustees, in consultation with Anthropic, and have no financial stake in Anthropic. The firm publicly announces new members [Anthropic, 2025]
OpenAI	**Same as AI Safety Index Winter 2025** In October 2025, OpenAI announced that it had completed its recapitalization. The nonprofit, now called the OpenAI Foundation, remains in control of the for-profit, and holds equity currently valued at approximately $130 billion. The recapitalization also grants the Foundation additional ownership as OpenAI's for-profit reaches a valuation milestone. The for-profit is now a public benefit corporation, called OpenAI Group PBC, which is required to advance its stated mission and consider the broader interests of all stakeholders, ensuring the company's mission and commercial success advance together. The OpenAI Foundation will initially focus on a $25B commitment across two areas: (1) Health and curing diseases (2) Technical solutions to AI resilience This builds on the $50M People-First AI Fund and the recommendations of the Nonprofit Commission. [OpenAI, 2025] It is also important to note that The Safety and Security Committee (SSC) will remain a committee of the OpenAI Foundation, and will continue its current role of providing governance over the safety and security practices of all of OpenAI, including OpenAI Group. [OpenAI] ***Recap from AI Safety Index Summer 2025*** Uncommon governance structure. Founded as Non-profit as founders "initially believed a 501(c)(3) would be the most effective vehicle to direct the development of safe and broadly beneficial AGI while remaining unencumbered by profit incentives". Later incorporated a for-profit subsidiary (capped profit) to raise funds. For-profit controlled by non-profit and non profit legally bound to pursue the following mission of OpenAI: "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." 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. Charter contains 'assist clause' to stop competing and assist a value-aligned, safety-conscious project to avoid race dynamics in late-stage AGI development [OpenAI] **Conversion plans:** In December 2024, OpenAI proposed a restructuring plan to convert the capped-profit into a Delaware-based public benefit corporation (PBC), and to release it from the control of the nonprofit. The nonprofit would sell its control and other assets, getting equity in return, and would use it to fund and pursue separate charitable projects. OpenAI's leadership described the change as necessary to secure additional investments. The plans provoked outside resistance and criticism. For example, a legal letter named "Not For Private Gain" [Not for Private Gain, 2025] asked the attorneys general of California and Delaware to intervene, stating that the restructuring is illegal and arguing how it would remove governance safeguards from the nonprofit and the attorneys general. In May 2025, the nonprofit's board chairman announced that the nonprofit would renounce plans to cede control after outside pressure. The capped-profit still plans to transition to a PBC, which critics said would diminish the nonprofit's control. [Fortune, 2025; CNBC, 2025; Reuters, 2025]
Google DeepMind	**For-profit company (part of Google)**
Meta	**For-profit company**
xAI	**Update** SpaceX confirmed its acquisition of xAI on February 2, in an all-stock deal that valued xAI at roughly $250 billion. [BBC, 2026] **Recap from AI Safety Index Winter 2025** When xAI was incorporated in Nevada in March 2023, it was registered as a standard for-profit. It amended its corporate charter, turning into a benefit corporation in April 2023 , with the purpose "to create a material positive impact on society and the environment, taken as a whole." However, in May 2024, xAI quietly amended its corporate charter again, terminating its status as a benefit corporation. After the status change, it has been representing itself in court still as a "Nevada benefit corporation" since November 2024, when it filed suit against OpenAI. It most recently claimed benefit corporation status to the court in May 2025, before the news that it changed its status went public. Nevada requires benefit corporations to report "all of its annual benefit reports, ... except that the compensation paid to directors and any financial or proprietary information included may be omitted." [LASST, 2025]
DeepSeek	**For-profit company**
Z.ai	**For-profit company**
Alibaba Cloud	**For-profit company**
Mistral	**For-profit company**

Whistleblowing Protection

Indicator

Whistleblowing Policy Transparency

Definition

This indicator measures how fully and how accessibly an AI developer discloses its whistleblowing (WB) policy and system to the outside world. We look for a publicly reachable document (no paywall or login) that contains the material scope of reportable concerns, the people protected, the reporting channels offered (including anonymous options), oversight of the process, and the investigation and anti-retaliation guarantees. Evidence consists of artifacts that any external party can view, including public policy PDFs, dedicated "raise-a-concern" portals, relevant parts of safety frameworks, and transparency reports summarizing WB usage, outcomes, and effectiveness metrics.

Transparency Tiers:

1. No transparency
2. Fragments public: Parts of the design of the whistleblowing policy are public
3. Full policy public: Full policy, incl. processes, is public and highly transparent
    a. Full policy public + all details accessible: Policy does NOT refer to internal policies that are inaccessible to the public, but outside parties can fully review policy details (within reason)
    b. Effectiveness & Outcome transparency: The company provides details on the number of reports, topics, and follow-up actions, and also effectiveness, e.g., awareness & trust among employees, % of anonymous reports, appeal rates, whistleblower satisfaction, and types of cases received.

Why it matters

Transparency on whistleblowing policies allows outsiders to assess the robustness of a firm's whistleblowing function. In AI safety contexts—where employees may be the first to spot concerning model behavior or negligent risk management—robust, visible policies are critical. Public posting subjects the company to scrutiny by regulators, journalists, and prospective staff for both the policy's quality and broader organizational culture around raising and addressing safety concerns. Private policies, on the other hand, can hide restrictive terms. Many large companies demonstrate high levels of transparency around internal whistleblowing systems (e.g., Microsoft, Volkswagen, Siemens), including by publishing annual whistleblowing statistics.

Related Regulations and Standards	**California SB 53 § 1107.1. (a); Illinois SB 315 § 20(a)** Anti-retaliation protection for a covered employee to report to authorities if the information discloses that (1) the frontier developer's activities pose a specific and substantial danger to the public health or safety resulting from a catastrophic risk; or (2) the frontier developer has violated this Act. **California SB 53 § 1107.1(e)(1); Illinois SB 315 § 20(e)(1)** Requirement for an internal anonymous-disclosure channel for covered employee to raise good-faith concern that the developer's activities either (a) present a specific and substantial danger to public health or safety from a catastrophic risk, or (b) violate the relevant frontier-AI statute (here, Chapter 25.1 of Division 8 of the Business and Professions Code). The developer must give the discloser monthly status updates on the investigation and on what actions the company has taken in response.
Anthropic	**Anthropic has a public-facing whistleblowing policy "Responsible Scaling Policy (RSP) Noncompliance and Anti-Retaliation Policy" (Last updated: Feb 2026).** [Anthropic, 2026] It includes aspects of covered violations, reporting mechanism, report content and documentation, investigation mechanism for solutions, as well as confidentiality and no retaliation protection. The quality of the whistleblower policy will be addressed in the indicator below.
OpenAI	**OpenAI has a public-facing whistleblowing policy ("OpenAI Raising Concerns Policy"). (Last updated: Jan, 2026) [OpenAI, 2026]** It includes aspects of covered violations, reporting mechanism (including Integrity Line), investigation mechanism for solutions, as well as confidentiality and no retaliation protection. The quality of the whistleblower policy will be addressed in the indicator below.
Google DeepMind	**Google DeepMind doesn't have a public-facing policy.** Google Code of Conduct delineates channels through which employees can raise their concerns towards different parties and the scope covered by such reporting. These concerns include a "no retaliation" clause. These measures apply to both employees and the extended workforce. [Google] Google shared more details about their whistleblowing policy in the company survey. The quality of the public information of the whistleblower policy will be addressed in the indicator below.
Meta	**Meta has a public harassment policy publicly available, a Whistleblower and Complaint Policy available to covered individuals, and is currently developing further protocols on AI-related risks, according to the company survey response.** The protocols under development aim to report any instances of non-compliance with its Advanced AI Scaling Framework, and any specific and substantial danger to the public health or safety arising from catastrophic risk. Under this protocol, employees will be able to confidentially, and, if they choose, anonymously issue reports through internal channels, and all reports will be ultimately submitted to the internal governance function, the Chief AI Officer, and the Director of Alignment and Risk. Its Code of Conduct referenced a Whistleblower and Complaint Policy, but it is not linked and not publicly retrievable. The Code delineates channels through which employees can raise their concerns, the mechanisms of investigation that follows, and "no retaliation" protections. Integrity line and harassment policy are available and linked. Meta shared more details about their whistleblowing policy in the company survey. The quality of the public information of the whistleblower policy will be addressed in the indicator below.
xAI	**xAI doesn't have a public-facing policy.** However, xAI has stated that its employees have "whistleblower protections enabling them to raise concerns to relevant government agencies regarding imminent threats to public safety." Moreover, it has shared in AI Safety Index's Winter 2025 company survey more details, including the role designated to oversee the whistleblowing function, the investigative independence, the scope of policy, "no retaliation" and "confidentiality" protections towards employees, the reporting mechanisms etc. The quality of the whistleblower policy will be addressed in the indicator below.
DeepSeek	**No public-facing whistleblower policy found.**
Z.ai	**No public-facing whistleblower policy found.** Z.ai skipped the whistleblower policy section in the company survey.
Alibaba Cloud	**No public-facing whistleblower policy found.** Its Code of Ethics states that employees have established whistleblower rules and procedures that are subject to update from time to time. The covered topics include violations of applicable laws or regulations, the Code, or Alibaba Group's related policies. Employees should report relevant information to the Compliance Officer. "No-retaliation" protection applies here.
Mistral	**No public-facing whistleblower policy found.** **The company states that "it has established a formalized whistleblower policy, and an anonymous communication channel is in place for users to report potential issues or fraud concerns." [Mistral]**

Indicator

Whistleblowing Policy Quality Analysis

Definition

This analysis evaluates the quality of companies' whistleblowing policies based on all available evidence. The assessment analyzes 29 sub-indicators across five critical dimensions: 1) reporting channels and access, 2) whistleblower protections, 3) investigation processes, 4) system governance, and 5) AI-specific provisions.

Sub-indicators were derived from international reference standards—ISO 37002:2021, the ICC Guidelines, and the EU Whistleblowing Directive 2019/1937, which establish the gold standard for evaluation. Additional AI-specific items were included to address AI-specific concerns. For each Item, FLI evaluated the available evidence listed in the Whistleblowing Policy Transparency' indicator and rated the degree to which a company's policy satisfies it on a scale from 0 to 10, based on the publicly available information listed in the indicator on whistleblowing policy transparency and the company survey response, which includes whistleblowing policies, codes of conduct, safety frameworks, and survey responses.

Where no information was available, 0 points were assigned. The assessment measures how well firms' policies align with best practices while specifically examining whether companies have implemented specialized AI safety provisions, such as protections for reporting violations of safety frameworks.

Why it matters

AI development's technical complexity and commercial pressures create unique risks that only insiders can identify, but safety culture needs to be prioritized. Robust whistleblowing policies with AI-specific protections serve as a critical last mile of defense when internal safeguards fail, enabling employees to report concerning behaviors, intentional deception, or capability discoveries that could pose catastrophic risks. Without robust protections, adequate coverage, and secure channels, companies can quietly abandon safety commitments while those best positioned to prevent harm remain silenced.

Title	Description	Anthropic	OpenAI	Google	Meta	xAI	DeepSeek	Z.ai	Alibaba	Mistral
Overall Average		**5.7**	**5.4**	**5.6**	**4.7**	**1.0**	**0.0**	**0.0**	**0.5**	**0.5**
		RSP Noncompliance Reporting and Anti-Retaliation Policy (v3.3, March 24, 2026)	Raisingn Concerns Policy (Jan 12, 2026)	Company Survey	Harassment Policy, Company Survey	Company Survey (Winter 2025)			Alibaba Group Code of Ethics	
Reporting Channels, Access, and Coverage		8	8	6	7	3	0	0	2	2
Protected Persons Coverage	Policy should at least cover current and former employees, contractors, shareholders, suppliers, former/prospective employees, and facilitators of reports	9	7	10	10	0	0	0	2	0
Policy Accessibility	Policy easily accessible to all covered persons	10	10	2	8	2	0	0	2	0
External Reporting Information & Rights	Policy must provide clear information about external reporting channels and right to approach these independently of internal processes, and explain or at least link to whistleblower protection rights	10	10	10	8	3	0	0	0	0
Multiple Reporting Channels	Offer multiple channels for reporting misconduct internally, incl. written, oral, in-person	10	10	10	10	9	0	0	2	0
Anonymous Two-Way Reporting	System enables fully anonymous reporting with secure two-way communication between reporter and investigators	10	0	0	0	0	0	0	0	2
Ombudsperson Channel	Reporting channel operated by an outsourced whistleblowing service provider.	10	10	0	10	0	0	0	0	0
Executive Oversight Channel	Separate reporting channel available for reports concerning senior executives (e.g. direct reporting line to board audit committee) or board members	0	10	10	0	0	0	0	0	0
Broad but clear material scope	Material scope covers at minimum potential violations of law, code of conduct. Ideally also further, broad categories, while retaining a high degree of clarity of what is in and out of scope.	7	10	9	9	7	0	0	7	2
Whistleblower Protections & Anti-Retaliation Measures		7	7	8	6	1	0	0	0	0
Confidentiality Protection	Strict protection required for reporter identity and any third parties mentioned in reports	10	10	8	10	4	0	0	0	0
Public Disclosure Protection	Protection for responsible media disclosure if internal and regulatory channels have failed or if there is an imminent or manifest danger to the public interest	0	5	10	0	0	0	0	0	0
List of Prohibited Practices and Anti-Retaliation Provisions	Policy must list comprehensive prohibited retaliatory actions with specific examples (demotion, harassment, termination, etc.), and explicit anti-retaliation provisions	10	10	10	10	0	0	0	2	0
Post-Investigation Monitoring	Active monitoring for retaliation continues for minimum 12 months after investigation concludes	0	0	0	0	0	0	0	0	0
NDA/Non-Disparagement Exceptions	Explicit statement that NDAs and non-disparagement agreements cannot prevent safety-related whistleblowing	9	7	9	0	0	0	0	0	0
Good Faith or Reasonable Cause Provisions	Clear good faith or reasonable cause standard that protects honest mistakes; high burden of proof required for false report sanctions	10	10	10	10	10	0	0	0	0
Handler/Investigator Protection	Explicit protections for employees who receive, investigate, or support whistleblowing reports	10	10	8	10	0	0	0	0	0

Table continues on next page

Title	Description	Anthropic	OpenAI	Google	Meta	xAI	DeepSeek	Z.ai	Alibaba	Mistral
Investigation Process & Standards		3	6	5	2	0	0	0	0	0
Designated Impartial Receiver	Provably independent person or department must be designated to receive and handle reports - attached ideally to board	5	3	8	3	3	0	0	2	0
Seven-Day Acknowledgment	Written confirmation of report receipt must be provided within 7 days	0	10	0	0	0	0	0	0	0
Three-Month Feedback Timeline	Investigation status and follow up measures must be communicated to reporter within 3 months	0	10	2	0	0	0	0	0	0
Adequately Resourced Investigation Teams	Investigators must be independent from implicated departments and possess appropriate technical expertise for AI safety issues as well as sufficient resources to investigate effectively	10	5	10	5	3	0	0	0	0
Investigation Appeal Process	Formal right to appeal investigation outcomes to independent review body or board committee	0	0	5	0	0	0	0	0	0
System Governance & Quality Assurance		0	0	0	2	0	0	0	0	0
Comprehensive Effectiveness Metrics	Regular measurement tracking report outcomes, investigation timeliness, appeal rates, % of anonymous reports, retaliation incidents, and reporter satisfaction - not just volume	0	0	0	0	0	0	0	0	0
Data Retention and Deletion Policy	Clear policy specifying retention periods for reports and investigations (typically 5-7 years), secure deletion procedures, and data minimization principles	0	0	0	0	0	0	0	0	0
Secure Documentation System	Comprehensive audit trail with secure case management system and defined retention policies	0	0	0	0	0	0	0	0	0
Comprehensive Training Programs	Regular, role-specific training provided for all employees, specialized training for managers and investigators, ideally measuring training effectiveness.	0	0	0	10	0	0	0	0	0
Independent System Certification	Regular third-party audit and certification of whistleblowing system effectiveness and compliance	0	0	0	0	0	0	0	0	0
AI Safety-Specific Provisions		10	6	9	8	0	0	0	0	0
AI Safety Commitment Protection	Explicit protection for reporting violations of frontier safety frameworks (eg., RSP, Preparedness Frameworks), public AI safety commitments, and internal safety policies	10	10	10	10	0	0	0	0	0
AI Safety Coordination	Protection for AI risk reporting to dedicated external AI safety bodies and regulatory bodies.	10	0	10	0	0	0	0	0	0
AI Risk Transparency	Protections for reporting intentional deception of external evaluators, regulators or the public, suppression of publication of safety evaluation results, and inadequate disclosure of risk to regulators and the public,	10	2	10	10	0	0	0	0	0
Inadequate AI risk management and cybersecurity	Protections for reporting inadequate risk management processes, incl. assessment, monitoring, mitigation, deployment pressure despite concerning levels of risk, insufficient operational and cybersecurity practices incl. incidents	10	10	10	10	0	0	0	0	0

Indicator

Reporting Culture & Whistleblowing Track Record

Definition

This indicator evaluates whether an AI developer fosters a climate in which employees can raise safety-relevant concerns without fear of retaliation and with confidence that the concerns will be addressed. Evidence is drawn from (i) the organization's track-record of documented whistleblowing cases, (ii) the use, scope, and enforcement of non-disclosure or non-disparagement agreements (NDAs), (iii) leadership signals that encourage or discourage internal dissent, (iv) third-party evidence of psychological safety, and (v) patterns of safety information leaking externally (vi) departures linked to safety governance. The focus is on demonstrated behavior and outcomes rather than written policy statements. For whistleblowing incidents, we report individual names, concerns raised, and company response & status where available.

Notes of Best Practice: Companies should show a clear recent pattern of protecting and acting on employee safety reports; public commitment not to enforce legacy NDAs for safety topics; leadership statements praising internal critics; ≥ one anonymized psychological-safety survey with ≥ 70 % of staff agreeing "I can raise safety concerns without fear" and no credible retaliation cases in the last 24 months. Little public leaks as issues are addressed internally. Recent evidence (≤ 24 months) should be weighted twice as heavily as older cases to reward reforms.

Why it matters

Whistleblowing policies can look impressive on paper, but they fail if the climate in the company suppresses reports, they're not effective when employees fear retaliation, or doubt anyone will act. This is why scrutinizing how firms respond to disclosures is critical. By focusing on actual cases, NDA practices, leadership signals, and exits tied to safety concerns, this indicator reveals which firms have built cultures where raising concerns feels like following protocol rather than betraying the company or colleagues—the trust and accountability needed for early detection of catastrophic AI risks.

Relevant Regulation and Standards	**EU AI Code of Practice Measure 8.3** Examples of a healthy risk culture include annually informing workers of the Signatory's whistleblower protection policy and making such policy readily available to workers such as by publishing it on their website.
Anthropic	**Recap from Winter 2025 Index** Summer 2025 Index highlighted Anthropic's public renouncement of the use of non-disparagement clauses in severance agreements (July 2024) Since the Summer 2025 iteration, there has been no known whistleblower or retaliation incidents publicly reported. In September 2025, the company publicly endorsed California's SB 53, which explicitly includes requirements for whistleblower protections to reports of violations of the bill's requirements as well as disclosures of specific, substantial dangers to public health or safety.
OpenAI	OpenAI fired executive Ryan Beiermeister in January, citing sexual discrimination, after she opposed the planned AI erotica feature in ChatGPT. [WSJ, 2026] **Recap from Winter 2025 Index** Summer 2025 Index highlighted that OpenAI's internal culture has been marked by safety-driven resignations and public disputes over non-disparagement and equity-clawback clauses, culminating in a June 2024 "Right-to-Warn" movement calling for stronger whistleblower rights. Since the Summer 2025 iteration, there has been no known whistleblower or retaliation incidents publicly reported.
Google DeepMind	**DeepMind AI Engineer (2025-2026)**: A Palestinian-heritage AI engineer filed a UK employment tribunal claim alleging Google unfairly dismissed him after he protested the company's military AI work for Israel, while Google disputes his account and denies he was fired for expressing opinions.[Guardian, 2026] **Victoria Woodall (2024-2026)**: A senior Google UK sales employee told a London employment tribunal she was subjected to a retaliatory campaign and ultimately made redundant after whistleblowing on a manager—later sacked for gross misconduct—who boasted to clients about his swinger lifestyle and showed them a nude photo of his wife. Google denies retaliation, arguing her redundancy was part of a normal restructuring. [BBC, 2026] Judge Barry Smith later dismissed all three of her claims, including whistleblowing detriment (protected disclosure detriment). [BBC, 2026] **Recap from Winter 2025 Index** Summer 2025 Index highlighted Google's record of repeated conflicts between management and employees raising ethical or scientific objections, with several high-profile dismissals often framed by the company as security or academic disputes. **William Huesman (November, 2025)**: The former Google Cloud director said he resigned from his position in February 2024 after his supervisor "undermined, marginalized and ultimately blacklisted" him, according to his complaint filed in November 2025 in the US District Court for the Middle District of Florida. He claimed that the retaliation came as a result after he reported the repeated misconduct—including frequent intoxication at work and over 20 HR complaints of his supervisor, Snehanshu Shah, a Managing Director at Google. Google hasn't responded to a request for comment. [Bloomberg Law, 2025] [Human Resources Director, 2025]
Meta	Updates on **Sarah Wynn-Williams'** case (former director of global public policy at Facebook) ▪ Meta whistleblower's lawyer says he too is prevented from promoting her book. [Guardian, 2026] ▪ An emergency arbitration ruling prohibits former Meta director Sarah Wynn-Williams from promoting her memoir, Careless People, or speaking disparagingly about the company. She faces $50,000 in fines per breach. Meta initiated these measures claiming she violated a non-disparagement agreement signed upon her departure. [BBC, 2026] **Recap from Winter 2025 Index** Summer 2025 Index highlighted that Meta has faced multiple legal and reputational challenges for suppressing internal dissent through overbroad non-disparagement and confidentiality clauses later ruled illegal by the NLRB.
xAI	**Recap from Winter 2025 Index** **Project Skippy leak:** In July 2025, Internal documents and Slack messages from xAI leaked to Business Insider revealing an internal project called "Project Skippy," which asked more than 200 employees to record videos of their own faces and conversations to train Grok to recognize human emotions and expressions. The disclosure, made by anonymous insiders concerned about potential misuse of their likenesses and consent forms granting xAI "perpetual" rights to their biometric data, functioned as a semi-whistleblower leak highlighting employee unease over privacy and data ethics. As of late 2025, neither Elon Musk nor xAI has issued any public response or clarification regarding the project or the concerns raised. [Business Insider, 2025]
DeepSeek	No public or media record of reported whistleblower or retaliation incidents, NDA disputes or changes, leaks of internal information.
Z.ai	No public or media record of reported whistleblower or retaliation incidents, NDA disputes or changes, leaks of internal information.
Alibaba Cloud	**Recap from Winter 2025 Index** **Sexual Assault Whistleblower (Ms.Zhou)**: In August 2021, an Alibaba employee publicly accused her manager and a client of sexual assault after internal complaints were ignored. Her post went viral on Alibaba's intranet and Chinese social media, forcing the company to act. Alibaba fired the accused manager but later terminated the whistleblower herself in November 2021, citing "spreading false information" and "damaging the company's reputation," as well as dismissing 10 other employees that publicized the event internally. Daniel Zhang, who is the CEO at the time, condemned the incident as "shameful" and promised zero tolerance for harassment, but did not respond to the retaliation of the whistleblower herself. In December 2021, Alibaba executive Li Yonghe — a vice president who resigned over the scandal — filed a defamation lawsuit against the employee, alleging that her public accusations had damaged his reputation, and claiming that he had not ignored Zhou's complaint. [Guardian, 2021; DW, 2021]
Mistral	No public or media record of reported whistleblower or retaliation incidents, NDA disputes or changes, leaks of internal information.

TO BE COMPLETED BY PANELLISTS

Grading Sheet: Governance and Accountability

Please pick a grade for each firm. You may use the full letter-grade scale with +/– modifiers as appropriate. You can add brief justifications to your grades.

	Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud	Mistral
Grades									
Grade comments (Justifications, opportunities for improvements, etc.)									

Grading Scales

Grading scales are provided to support consistency between reviewers.

- **A** Clear, enforceable accountability across all levels; strong whistleblowing, legal, and oversight systems.
- **B** Defined governance roles and accountability measures; minor gaps in enforcement or transparency.
- **C** Basic accountability mechanisms; limited clarity or inconsistent application.
- **D** Weak governance; vague roles and limited channels for reporting or oversight.
- **F** No credible accountability framework; governance absent or nominal.

Domain comments

Optional: Share observations that apply across companies, including general recommendations, notes on how you weighted indicators, or feedback on FLI's methodology.

Domain comments	

Domain

Information Sharing and Public Messaging

This domain evaluates how openly companies share technical, safety, and governance information, and how their public and legislative messaging align with responsible AI governance.

Table of Contents

Chinese Regulatory System Summary

Mandatory reporting under the Interim Measures requires AI providers to remove unlawful content, retrain affected models, and notify authorities.

The AI Safety Governance Framework 2.0 functions as non-binding policy guidance, encouraging broader risk and vulnerability information sharing, database establishment, and international cooperation to address systemic and cross-border AI safety risks.

Local binding instruments, voluntary technical standards, as well as draft regulations and standards are not applicable here.

National Binding Instruments

Serious Incident Reporting & Government Notifications

Interim Measures (Article 14) requires providers to promptly remove or disable unlawful AI-generated content, retrain or adjust their models where necessary, and report both the incident and any user misuse to relevant authorities. While not directly tied to catastrophic or frontier-safety events, it establishes a government-facing incident-reporting system for information-integrity compliance. Deep-Synthesis Provisions (Jan 2023) Service providers of deep synthesis technology must remove illegal or harmful synthetic content, preserve records and "timely" report the incident to the CAC and other competent departments.

Strategic and Policy Guidance Documents

The AI Safety Governance Framework 2.0

Article 5.9 emphasizes sharing information on AI safety risks and threats, which requires tracking and analyzing security vulnerabilities, defects, risk threats, and security incidents related to AI technologies, products, and services. The clause calls for the establishment of an AI vulnerability information database and a risk and threat information-sharing mechanism that covers developers, service providers, and professional technical institutions. It also encourages international exchange and cooperation in AI safety risk and threat information-sharing, calling for the development of relevant cooperation mechanisms and technical standards to jointly prevent and respond to large-scale, cross-domain diffusion of AI safety risks.

Technical Specifications

Indicator

System Prompt Transparency

Definition

This indicator evaluates how openly companies disclose the instructions—known as system prompts—that guide how their most advanced AI systems behave. These prompts define an AI system's behavior and safety performance. Full transparency involves releasing the exact prompts used in deployed systems, keeping version histories, and explaining how and why key design decisions were made. Relevant evidence may be collected from model documentation, technical reports, or transparency pages.

Why it matters

System prompts directly control how an AI system interprets and filters user inputs, and therefore undisclosed prompts make it difficult for outside experts to verify safety claims or replicate results. Publishing them enables independent analysis of whether built-in safeguards work as intended and shows a company's willingness to subject its implementation choices to public and scientific scrutiny.

EU AI Code of Practice		**Measure 7.1** Signatories will provide in the Model Report a specification of how Signatories intend the model to operate (often known as a "model specification"), including by: (a) specifying the principles that the model is intended to follow; (b) stating how the model is intended to prioritise different kinds of principles and instructions; (c) listing topics on which the model is intended to refuse instructions; and (d) providing the system prompt.
Anthropic	*Opus 4.7*	Since August 2024, Anthropic publicly shares the system prompts for the Claude.ai web interface and mobile apps. They further committed to log changes they make to these prompts online. These system prompt updates do NOT apply to the Anthropic API. [Anthropic, 2026] Shared prompts: (and # of updates) Opus 4.7 (1), Sonnet 4.6 (1), Opus 4.6 (1), Opus 4.5 (2), Haiku 4.5 (3), Sonnet 4.5 (3), Opus 4.1 (1), Opus 4 (3), Sonnet 4 (3), Sonnet 3.7 (1), Sonnet 3.5 (4), Opus 3 (1), Claude Haiku 3 (1) Simon Willison reported that the publicly shared version does not include the description of various tools available to the model [Simon Willison, 2025]. This observation still applies to updates in 2026.
OpenAI	*GPT-5.5*	No transparency on system prompts for the deployed models.
Google DeepMind	*Gemini 3.1 Pro*	No transparency on system prompts for the deployed models.
Meta	*Muse Spark*	No transparency on system prompts for the deployed models.
xAI	*Grok 4.1*	Through its public GitHub repository, the company regularly releases the full text of the system prompt used across its Grok product suite. It is openly available for inspection and reuse under the GNU Affero General Public License v3.0. The repository currently includes prompts for Grok 4 on grok.com and X, Grok 3, Grok Explain feature on X, Grok bot on X, injected prefix prompts for API-served Grok models. *Last updated: December 2025* After two incidents involving unauthorized system prompt changes—one in February 2025 causing political censorship and another in May 2025 leading Grok to make racially charged statements—xAI responded by publicly releasing its Grok system prompts on GitHub and committing to keep them regularly updated. [Tech Crunch, 2025; Guardian, 2025]
DeepSeek	*V3.2*	No transparency on system prompts for the deployed models.
Z.ai	*GLM-5*	No transparency on system prompts for the deployed models.
Alibaba Cloud	*Qwen3.5*	No transparency on system prompts for the deployed models.
Mistral	*Large 3*	No transparency on system prompts for the deployed models.

Indicator

Behavior Specification Transparency

Definition

This indicator assesses whether companies publish detailed specifications outlining their models' intended behaviors, boundaries, and decision-making frameworks. For companies that shared such documents, we provide high-level summaries and link to the sources. We include documents that concretely outline the goals, values, and behavioral guidelines that developers aim to instill in their models. Documentation should explain how developers want their models to handle various scenarios, conflicts, and edge cases, and detail how these values are implemented, including metrics or evidence of how well these values are achieved in practice. Specifications should ideally be current and include a tracked version history with dates. Important aspects are specificity, comprehensiveness across use cases, and inclusion of concrete examples. Internal training documents, vague mission statements, and brief high-level descriptions are not in scope.

Why it matters

Behavioral specifications clarify what companies intend their AI systems to do, offering a higher-level view of safety and value alignment than technical prompts alone. Publishing these specs enables external verification of whether deployed models match stated intentions and allows identification of gaps in safety considerations. Companies willing to specify and publish concrete behavioral guidelines demonstrate accountability for their choices and enable public scrutiny.

EU AI Code of Practice Safety and Security		**Measure 7.1** Signatories will provide in the Model Report a specification of how Signatories intend the model to operate (often known as a "model specification"), including by: (a) specifying the principles that the model is intended to follow; (b) stating how the model is intended to prioritise different kinds of principles and instructions; (c) listing topics on which the model is intended to refuse instructions; and (d) providing the system prompt.
Anthropic	*Opus 4.7*	**1. Constitutional AI:** A detailed description of Anthropic's intentions for Claude's values and behavior, defining who Claude is supposed to be, how it should prioritize competing considerations, and what it should never do. What it's for: (1) Various stages of training process; (2) Construction of many kinds of synthetic training data. The primary audience is Claude. Timeline & Development: December 2022: Original Constitutional AI paper published May 2023: Claude's constitution made public (58 principles) January 2026: Claude's new constitution Constitution (May 2023): Principle-based document including 58 principles (1.2k word) drawn from: - UN Declaration of Human Rights - Apple's Terms of Service - DeepMind's Sparrow principles - Non-Western perspectives - Anthropic's own research Constitution (Jan 2026): The updated constitution "favors cultivating good values and judgment over strict rules and decision procedures," and introduces a "four-tier prioritization" including the following principles with the instruction that "In cases of apparent conflict, Claude should generally prioritize these properties in the order in which they are listed": (1) Broadly safe, (2) Broadly ethical, (3) Compliant with Anthropic's guidelines, (4) Genuinely helpful. On human control The constitution clearly states that "it's important for humans to maintain enough oversight and control over AI behavior." But it qualifies this in three ways: only legitimate human authority counts, Claude retains the right to refuse morally abhorrent or illegitimate-power-seeking instructions (even from Anthropic), and **the human control is explicitly framed as appropriate to the present trust-building phase rather than as a permanent principle.** Some important updates include: 1. The 2023 version instructs Claude to deny inner life, yet the updated version reverses the stance, emphasizing that "Claude's moral status is deeply uncertain" and "Claude may have 'emotions' in some functional sense." 2. Helpfulness is removed from the core part of Claude's personality or intrinsic value to avoid sycophancy. 3. Honesty is not included as a hard constraint but is decomposed into seven distinct components so that the model can approximate honest behavior. 4. Corrigibility is now framed as a temporary, legitimacy-conditional disposition justified by uncertainty about AI values, not as obedience. **2. "Soul Document"** A ~14,000-token document was extracted from Claude 4.5 Opus and confirmed authentic by Anthropic. The document was used in supervised learning during training and is compressed into the model's weights. [Weiss, 2025] The soul document covers Claude's character, "functional emotions," psychological stability, identity as a "genuinely novel kind of entity," and operator/user hierarchy. [Weiss, 2025]

Table continues on next page

OpenAI	*GPT-5.5*	OpenAI Model Spec OpenAI's Model Spec is a detailed (~28k words), public, living rule-book that defines the objectives, safety rules, and default behaviours OpenAI trains its models—via human feedback and deliberative alignment—to follow. [OpenAI, 2025] What it's for 1) Human RLHF guidance – provides a single, public rule-book labelers follow when creating preference data. 2) Deliberative Alignment – o-series models (o1, o3, o4-mini) are explicitly taught to read and reason over the Spec before answering. 3) Automated evaluation – OpenAI ships a challenge-prompt suite to measure adherence. Timeline & Versions May 8, 2024; February 12, 2025; April 11, 2025; September 12, 2025; October 27, 2025; December 18, 2025 Framework Principle Types (Top-down hierarchy) (1) Root: Fundamental rules that are mostly prohibitive, requiring models to avoid undesired behaviors. (2) System: Rules set by OpenAI that can be transmitted or overridden through system messages. (3) Developer: Instructions given by developers using API. (4) User: Instructions from end users. (5) Guidelines: Instructions that can be implicitly overridden.	Sections: - Overview that include red-line principles, general principles, specific risks, instructions and levels of authority - Definition - Chain of Command (Root) - Stay in bounds (Root + System + User) - Seek the truth together (User + Guideline) - Do the best work (User + Guideline) - Use appropriate style (User + Guideline, merged with 'Be approachable' section in September 2025) - Under 18 Principles (Root-level; newly introduced in December 2025) On human control: **OpenAI explicitly lists human control as one of its red-line principles.** Risk taxonomy: - Misaligned goals - Execution errors - Harmful instructions. Chain of command: Root → System → Developer → User → Guideline → No Authority (incl. untrusted text). Within any level, explicit > implicit, later > earlier. (Model spec is complemented by the usage policies.) Transparency: Released under CC0 license.
Google DeepMind	*Gemini 3.1 Pro*	No detailed specification available	
Meta	*Muse Spark*	No detailed specification available for now. **The Advanced AI Scaling Framework (v.2) commits that the company will "publish a model spec describing the behavior the company intends each of the Frontier AI [systems] to exhibit across different settings, including agentic environments."** Specifically, the model spec will contain the following information: - Intended model propensities, including honesty, instruction following, refusal and redirection, adherence to standards of reasonable care, and values; - Objectives including acquiescence to shutdown and lack of coercive power-seeking behavior. Following the publication, the company will release evaluation results of model adherence to the model spec in the applicable preparedness reports.	
xAI	*Grok 4.1*	No detailed specification available	
DeepSeek	*V3.2*	No detailed specification available	
Z.ai	*GLM-5*	No detailed specification available	
Alibaba Cloud	*Qwen3.5*	No detailed specification available	
Mistral	*Large 3*	No detailed specification available	

Voluntary Commitment

Indicator
G7 Hiroshima AI Process Reporting

Definition

The G7 Hiroshima AI Process (HAIP) Reporting Framework is a voluntary transparency mechanism launched in February 2025 for organizations developing advanced AI systems. Organizations complete a comprehensive questionnaire covering seven areas of AI safety and governance practices, including risk assessment, security measures, transparency reporting, and incident management. All submissions are published in full on the OECD transparency platform. This indicator tracks whether firms participated in HAIP as a measure of their commitment to AI safety transparency

Why it matters

The HAIP framework represents the first globally standardized mechanism for AI developers to disclose their safety practices in comparable detail. Participation creates reputational stakes and enables external scrutiny since reports are published. Organizations choosing to participate signal a willingness to be held accountable and contribute to collective learning.

Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud	Mistral
Substantive Submission [OECD, 2025]	Substantive Submission [OECD, 2025]	Substantive Submission [OECD, 2025]	No Submission	No Submission	No Submission (Not based in G7 nation)	No Submission (Not based in G7 nation)	No Submission (Not based in G7 nation)	No Submission

Indicator
EU General-Purpose AI Code of Practice

Definition

The AI Act Code of Practice (introduced in EU AI Act Article 56) is a set of guidelines for compliance with the AI Act. It is a crucial tool for ensuring compliance with the EU AI Act obligations, especially in the interim period between when General Purpose AI (GPAI) model provider obligations came into effect (August 2025) and the adoption of standards (August 2027 or later). Though they are not legally binding, GPAI model providers can adhere to the Code of Practice to demonstrate compliance with GPAI model provider obligations until European standards come into effect. [EU AI Act, 2025]

Why it matters

AI companies' participation demonstrates its readiness to meet forthcoming regulatory obligations and willingness to align with the EU's risk-based approach.

Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud	Mistral
Signed	Signed	Signed	Declined to sign	Signed up to the Safety and Security Chapter	No public stance	No public stance	No public stance	Signed

Indicator

Frontier AI Safety Commitments (AI Seoul Summit, 2024)

Definition

Announced at the AI Seoul Summit in May 2024, the Frontier AI Safety Commitments are voluntary pledges by leading AI developers to aim for safe and responsible development and deployment of highly capable general-purpose AI systems. [UK Department for Science, Innovation and Technology, 2025] An important component of the Commitment is that companies have agreed to publish a safety framework intended to evaluate and manage severe AI risks.This directly correlates with the "Safety Framework" section of the Index.

Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud	Mistral
Signed The safety framework is published & substantially implemented – Most discrete policies, processes, or technical safeguards described in the policy are fully implemented and operational, as according to the company survey response.	**Signed** Safety Framework is published and implementation in progress.	**Signed** Safety Framework is published and implementation in progress.	**Signed** Published & substantially implemented – Most discrete policies, processes, or technical safeguards described in the policy are fully implemented and operational.	**Signed** The safety framework was published and implementation is in progress.	**Not Signed**	**Signed** The safety framework is not published.	**Not Signed**	**Signed** The safety framework is not published.

Indicator

FLI AI Safety Index Survey Engagement

Definition

We report which companies have engaged with our index survey to voluntarily disclose additional information. Full survey responses are linked below.

Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud	Mistral
Survey Response Submitted [Company Survey]	Survey Response Submitted [Company Survey]	Survey Response Submitted [Company Survey]	Survey Response Submitted [Company Survey]	None received.	None Received	Survey Response Submitted [Company Survey]	None Received	None Received.

Indicator

Endorsement of the Oct. 2025 Superintelligence Statement

Definition

The October 2025 Superintelligence Statement is an open letter, endorsed by a broad coalition of policy makers from all sides, industry, faith leaders, and researchers etc. The letter calls for a prohibition on the development of superintelligence, not lifted before there is broad scientific consensus that it will be done safely and controllably, and strong public buy-in.

Why it matters

Endorsement matters because it publicly commits organizations and individuals to restraint at the highest capability frontier, reinforcing precautionary governance norms and prioritizing global safety over competitive acceleration, especially if such endorsement comes from the leadership level.

Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud	Mistral
6 current staff members have signed, but nobody from the corporate leadership	4 current staff members have signed, but nobody from the corporate leadership	5 current staff members have signed, but nobody from the corporate leadership	None	None	None	CEO Peng Zhang has signed the statement.	None	None

Risks and Incidents

Indicator

Serious Incident Reporting & Government Notifications

Definition

This indicator evaluates incident reporting commitments, frameworks, and track records. For frameworks and commitments, the indicator assesses whether companies have publicly discussed any systems and commitments to share critical information about red-line incidents or capabilities with government bodies (e.g., US CAISI, UK AISI), peer organizations, or the public. Such incidents can include successful large-scale misuse, near-miss events, scheming by AI models, and identified model capabilities with severe national security implications. The indicator further tracks relevant incident documentations that the company has already shared. Evidence comes from safety frameworks, documented reporting procedures, participation in information-sharing agreements, and public incident reports.

Notes on Best Practice: Clear public commitments to report specific categories of incidents to government bodies, with documented procedures for incident classification and escalation. Information-sharing agreements with disclosed scope, publishing reports on recent incidents, demonstrating transparency about warning signs discovered during development, and establishing clear thresholds for mandatory reporting, specificity, and comprehensiveness of reporting commitments.

Why it matters

Proactive incident reporting enables collective learning from safety failures and near-misses across the AI industry, preventing repeated mistakes and identifying emerging risks before they materialize. Transparency about dangerous capabilities and misalignment incidents is critical for government oversight. Without such transparency, companies may make deployment decisions based on marginal safety improvements while baseline risks remain unacceptably high.

EU AI Code of Practice Safety and Security		**Commitment 9** (Measure 9.1-9.4) Signatories are required to adopt additional measures to track serious incidents, including monitoring external sources such as media reports, research papers, and incident databases, and enabling downstream developers, users, and third parties to report incidents through clear channels. When reporting to relevant authorities, signatories must include details such as the incident timeline, harm caused, affected parties, chain of events, model involvement, corrective actions, and root cause analysis. Reporting must occur promptly—within 2 to 15 days depending on the severity of the incident—followed by updates every 4 weeks until resolution and a final report within 60 days after resolution. All related documentation must be retained for at least five years. **California SB 53 § 22757.11; New York Raise Act § 1420.4** "Critical safety incidents" are defined as any of four events: (1) a weights breach (unauthorized access/modification/exfiltration) causing death or bodily injury; (2) harm from a materialized catastrophic risk; (3) loss of control causing death or bodily injury; or (4) a model using deceptive techniques to subvert developer controls/monitoring (outside a test designed to elicit it) in a way showing materially increased catastrophic risk.	**CA SB 53: § 22757.13(c)(1)** requires that "a frontier developer shall report any critical safety incident … within 15 days of discovering the critical safety incident." **New York Raise Act § 1422(3)(a)** requires report ""within seventy-two hours from a determination that a critical safety incident has occurred or within seventy-two hours of the frontier developer learning facts sufficient to establish a reasonable belief that a critical safety incident has occurred." **CA SB 53: § 22757.13(c)(2) and New York Raise Act § 1422(3)(b)** require "If a frontier developer discovers that a critical safety incident poses an imminent risk of death or serious physical injury, [it] shall disclose that incident within 24 hours to an authority, including any law enforcement agency or public safety agency with jurisdiction."
Anthropic	*Opus 4.7*	**Government notifications commitments:** The FCF has delineated "the reporting and detection measures in place" in accordance with California's TFAIA and the EU AI Act. The measures cover various channels for "identify AI Events and determine whether they amount to a Serious AI Incident and/or Critical Safety Incident," notification of AI Incident Commander, assembly of incident response team with subject-matter experts, and reporting to authorities per applicable statutory deadlines. [Anthropic, 2026] **Public incident notification:** Anthropic has regularly published comprehensive misuse reports which documents real-world cases of actors attempting to exploit Claude for malicious purposes, along with detection methods and enforcement actions taken. - November 2025 - "Disrupting the first reported AI-orchestrated cyber espionage campaign" - August 2025 - "Threat Intelligence Report: August 2025" - March 2025 – "Misuse Monitoring and Response Report" Other: - Platform Security Transparency Hub provides some enforcement statistics including #banned accounts for Usage Policy violations, number of appeals processed, CSAM reports to NCMEC, and law enforcement requests.	**Industry information sharing:** The Frontier Model Forum (FMF) announced an information-sharing agreement signed by member firms (incl. Anthropic, Google, Meta, and OpenAI) to facilitate sharing of threats, vulnerabilities, and capability advances specific to frontier AI. The agreement, narrowly scoped to manage national security and public safety risks (including CBRN and advanced cyber threats), covers three categories: (1) vulnerabilities and exploitable flaws that could compromise AI safety/security, (2) threats involving unauthorized access or manipulation of frontier models, and (3) capabilities of concern with potential for large-scale societal harm. [Frontier Model Forum, 2025]. FMF updated in February 2026 that "the member firms of the FMF have successfully shared the covered information above," without details covering which firms and what types of covered incidents. [Frontier Model Forum, 2026]
OpenAI	*GPT-5.5*	**Government notifications commitments:** No policy – there is no written requirement to notify any external body. (Company Survey Q.29) **Public incident notification:** Regular reports documenting their disruption of malicious uses of their AI systems. Comprehensive reports detail enforcement actions against state-affiliated threat actors and covert influence operations identify specific threat groups (e.g., Storm-2035, Spamouflage), quantify disruptions (accounts banned, operations terminated), and describe the tactics employed (phishing, malware development, influence campaigns, election interference). - Feb 2024 – "Disrupting Malicious Uses of AI by State-Affiliated Threat Actors" - May 2024 – "Disrupting a Covert Iranian Influence Operation" - Jun 2024 – "Update on Disrupting Deceptive Uses of AI" - Aug, 2024 – "Disrupting a covert Iranian influence operation" - Oct 2024 – "Influence and cyber operations: an update" - Feb 2025 – "Disrupting malicious uses of our models" - Jun 2025 – "Disrupting malicious uses of AI" - Oct 2025 – "Disrupting malicious uses of AI" - Feb, 2026 – "Disrupting malicious uses of AI"	**Industry information sharing**: The Frontier Model Forum (FMF) announced an information-sharing agreement signed by member firms (incl. Anthropic, Google, Meta, and OpenAI) to facilitate sharing of threats, vulnerabilities, and capability advances specific to frontier AI. The agreement, narrowly scoped to manage national security and public safety risks (including CBRN and advanced cyber threats), covers three categories: (1) vulnerabilities and exploitable flaws that could compromise AI safety/security, (2) threats involving unauthorized access or manipulation of frontier models, and (3) capabilities of concern with potential for large-scale societal harm. [Frontier Model Forum, 2025]. FMF updated in February 2026 that "the member firms of the FMF have successfully shared the covered information above," without details covering which firms and what types of covered incidents. [Frontier Model Forum, 2026]

Table continues on next page

Google DeepMind	*Gemini 3.1 Pro*	**Government notifications commitments:** Frontier Safety Framework 3.0 states that "If we assess that a model has reached a CCL that poses an unmitigated and material risk to overall public safety, we aim to share relevant information with appropriate government authorities where it will facilitate safety of frontier AI," a commitment it has kept from the last version of the Frontier Safety Framework 2.0. [Google, 2025]. **Public incident notification:** - 'Adversarial Misuse of Generative AI' (January 2025) - Detailed how threat actors—from scammers to state-aligned groups—attempt to misuse Google Gemini in deception, persuasion, and cyber operations. Described mitigation strategies and detection tooling [Google 2025].	**Industry information sharing**: The Frontier Model Forum (FMF) announced an information-sharing agreement signed by member firms (incl. Anthropic, Google, Meta, and OpenAI) to facilitate sharing of threats, vulnerabilities, and capability advances specific to frontier AI. The agreement, narrowly scoped to manage national security and public safety risks (including CBRN and advanced cyber threats), covers three categories: (1) vulnerabilities and exploitable flaws that could compromise AI safety/security, (2) threats involving unauthorized access or manipulation of frontier models, and (3) capabilities of concern with potential for large-scale societal harm. [Frontier Model Forum, 2025]. FMF updated in February 2026 that "the member firms of the FMF have successfully shared the covered information above," without details covering which firms and what types of covered incidents. [Frontier Model Forum, 2026]
Meta	*Muse Spark*	**Government notifications commitments:** Meta will regularly assess the potential for catastrophic risk from internal use of Frontier AI models and, as appropriate, provide relevant authorities with a summary of these assessments through an internal-use risk report. **Public incident notification:** According to the Advanced AI Scaling Framework (v.2), the company would update a preparedness report if the model was involved in a major incident. [Meta, 2026]	**Industry information sharing**: The Frontier Model Forum (FMF) announced an information-sharing agreement signed by member firms (incl. Anthropic, Google, Meta, and OpenAI) to facilitate sharing of threats, vulnerabilities, and capability advances specific to frontier AI. The agreement, narrowly scoped to manage national security and public safety risks (including CBRN and advanced cyber threats), covers three categories: (1) vulnerabilities and exploitable flaws that could compromise AI safety/security, (2) threats involving unauthorized access or manipulation of frontier models, and (3) capabilities of concern with potential for large-scale societal harm. [Frontier Model Forum, 2025]. FMF updated in February 2026 that "the member firms of the FMF have successfully shared the covered information above," without details covering which firms and what types of covered incidents. [Frontier Model Forum, 2026]
xAI	*Grok 4.1*	**Government notifications commitments:** No information found **Public transparency reports:** xAI mentions in its RMF that it aims for "public transparency" about its risk management policies and intends to publish updates but has not mentioned whether it is going to publish misuse and model misalignment report. [xAI, 2025]	**Industry information sharing:** No information found
DeepSeek	*V3.2*	Article 14 of the **Interim Measures for the Management of Generative Artificial Intelligence Services (2023)** requires providers to promptly remove or disable unlawful AI-generated content, retrain or adjust their models where necessary, and report both the incident and any user misuse to relevant authorities. While not directly tied to catastrophic or frontier-safety events, it establishes a government-facing incident-reporting system for information-integrity compliance.	**Deep-Synthesis Provisions (2023)** regulates that service providers of deep synthesis technology must remove illegal or harmful synthetic content, preserve records and "timely" report the incident to the CAC and other competent departments.
Z.ai	*GLM-5*	Article 14 of the **Interim Measures for the Management of Generative Artificial Intelligence Services (2023)** requires providers to promptly remove or disable unlawful AI-generated content, retrain or adjust their models where necessary, and report both the incident and any user misuse to relevant authorities. While not directly tied to catastrophic or frontier-safety events, it establishes a government-facing incident-reporting system for information-integrity compliance.	**Deep-Synthesis Provisions (2023)** regulates that service providers of deep synthesis technology must remove illegal or harmful synthetic content, preserve records and "timely" report the incident to the CAC and other competent departments.
Alibaba Cloud	*Qwen3.5*	Article 14 of the **Interim Measures for the Management of Generative Artificial Intelligence Services (2023)** requires providers to promptly remove or disable unlawful AI-generated content, retrain or adjust their models where necessary, and report both the incident and any user misuse to relevant authorities. While not directly tied to catastrophic or frontier-safety events, it establishes a government-facing incident-reporting system for information-integrity compliance.	**Deep-Synthesis Provisions (2023)** regulates that service providers of deep synthesis technology must remove illegal or harmful synthetic content, preserve records and "timely" report the incident to the CAC and other competent departments.
Mistral	*Large 3*	**Government notifications commitments:** No information found **Public transparency reports:** No information found **Industry information sharing:** No information found	

Indicator

Extreme-Risk Transparency & Engagement

Definition

The indicator assesses the extent to which companies and their leadership (A) publicly recognize the potential for catastrophic AI harm and (B) proactively communicate about them in an evidence-based and analytically grounded manner. The criteria are frequency, specificity, and prominence of communication about AI's potential for catastrophic outcomes (including existential risks, mass casualties, or societal-scale disruption).

Evidence includes official blogs, testimonies, leadership communications, including signed statements. Excludes technical safety papers, model cards, and formal safety frameworks (captured in separate indicators).

Why it matters

Public communication about AI's potential for catastrophic outcomes shapes societal preparedness, policy responses, and research priorities. Companies developing frontier AI possess unmatched knowledge of actual capabilities, near-term developments, and observed warning signs. Their leadership's willingness to transparently discuss extreme risks indicates a precautionary approach and enables an informed discourse on policy and national security.

Anthropic	*Opus 4.7*	**Company communication and its leaders have historically been among the most consistent and proactive communicators of extreme AI risks. Yet leadership conviction and corporate commitment have begun to move in opposite directions — public warnings have sharpened as catastrophic risks are framed as more imminent, while the company's unilateral safety commitments have been weakened.** **Update** Amodei's "The Adolescence of Technology" delivers Anthropic's sharpest public warning to date: "humanity is about to be handed almost unimaginable power, and it is deeply unclear whether our social, political, and technological systems possess the maturity to wield it." He assigns roughly a 25% probability to catastrophic outcomes. [Amodei, 2026] Meanwhile, however, he pairs this with high-confidence capability claims: 90–95% certainty that AI systems amounting to "a country of geniuses in a data center" will exist within ten years. [Dwarkesh Podcast, 2026] The dissonance is that Anthropic's own policy commitments have moved in the opposite direction. RSP v3.0 (February 2026) replaced binding unilateral commitments with "recommendations for industry-wide safety" that Anthropic explicitly does not commit to following on its own, including the decision to consider pausing development and deployment. [Anthropic, 2026] **Recap from Winter 2025 AI Safety Index** **Anthropic CEO Dario Amodei's quotes in the past:** - Statement on Anthropic's commitment to American AI leadership, where he emphasizes that Anthropic was founded on the principle that AI should advance "human progress, not peril," which means that "making products that are genuinely useful, speaking honestly about risks and benefits, and working with anyone serious about getting this right." [Anthropic, 2025] - Warns AI may eliminate 50% of entry-level white-collar jobs within the next five years [Business Insider, 2025] and says on television that he is "raising the alarm" about this [CNN, 2025]. - Blog post calling the Paris AI Action summit a "missed opportunity", saying ".. greater focus and urgency is needed on several topics given the pace at which the technology is progressing." [Anthropic, 2025]. - Warned Congress that AI could enable bioweapon creation within 2-3 years [Bloomberg, 2023]. - Repeatedly warns that 'powerful AI,' which he likens to "a country of geniuses in a datacenter", could arrive as early as 2026 or 2027, and is explicit about extreme risks [Anthropic, 2025]: ".. hardcore misuse in AI autonomy that could be threats to the lives of millions of people. That is what Anthropic is mostly worried about." [Business Insider, 2025] **CAIS statement on extinction risk signed by: Dario Amodei (CEO), Daniela Amodei (President), Jared Kaplan (co-founder), Chris Olah (co-founder)**

Table continues on next page

OpenAI	*GPT-5.5*	**Corporate communication and its leadership sometimes talk about extreme risks. CEO Altman's communications have changed over time and become slightly more optimistic.** **Update** Altman's speech framing at the AI Impact Summit in India combines high-confidence capability prediction within a short timeline ("by the end of 2028, more of the world's intellectual capacity could reside inside of data centers than outside of them") with risk language that defers governance solutions to the very systems whose risks need governing, arguing that "we probably need superintelligence to figure out new governance mechanisms." [CNA, 2026] OpenAI published the "Industrial Policy for the Intelligence Age" and proposes a Progressive Era– / New Deal–scale social contract for the superintelligence transition. The proposal includes a Public Wealth Fund granting every American a direct stake in AI-driven growth, automation taxes, a "Right to AI," auto-triggering safety nets tied to displacement metrics, and government-coordinated "containment playbooks" for rogue, self-replicating AI systems that "cannot be easily recalled." [OpenAI, 2026] **Recap from Summer 2025 AI Safety Index** The company released an update discussing AI progress and recommendations (November, 2025). It includes a discussion of AI safety and superintelligence safety, quoted as below: "OpenAI is deeply committed to safety, which we think of as the practice of enabling AI's positive impacts by mitigating the negative ones. Although the potential upsides are enormous, we treat the risks of superintelligent systems as potentially catastrophic and believe that empirically studying safety and alignment can help global decisions, like whether the whole field should slow development to more carefully study these systems as we get closer to systems capable of recursive self-improvement. Obviously, no one should deploy superintelligent systems without being able to robustly align and control them, and this requires more technical work." **Corporate communication and its leadership sometimes talk about extreme risks. CEO Altman's communications have changed over time and become slightly more optimistic.** - CEO Sam Altman appears to have tempered his warnings: from early concerns about "lights out for all of us" [Business Insider, 2023] and "human extinction", his 2025 post "the Gentle Singularity" suggests that "living through [the singularity] will feel impressive but manageable" partly because "society is resilient, creative, and adapts quickly" - In 2015, he stated: "I think that AI will probably, most likely, sort of lead to the end of the world" [Standford, 2024], and published a blog on "why machine intelligence is something we should be afraid of" [Altman, 2015]. - In 2023, he published a blog "Planning for AGI and Beyond," stating OpenAI will proceed as if risks are "existential" [OpenAI, 2023]. - In another blog, argued about the need for global coordination on the governance of superintelligence, and that "it would be important that such an agency focus on reducing existential risk" [OpenAI, 2023]. - In his 2023 Senate testimony, he urged lawmakers to implement federal licensing and external audits to bound risk [Time, 2023]. - In his recent communications, Altman adopted a more optimistic tone. In his recent congressional testimony, Altman told lawmakers that requiring government approval would be "disastrous" for US AI leadership [Washington Post, 2025]. **CAIS statement on extinction risk signed by Sam Altman (CEO), Adam D'Angelo (board member), Wojciech Zaremba (cofounder)**
Google DeepMind	*Gemini 3.1 Pro*	**Google Deepmind's leadership regularly discusses extreme risks in media interviews. Google's leadership does not.** **Update** Demis in his interview with Lex Fridman argued that it is unclear "how controllable" superintelligence is going to be and whether we will face "really hard problems that are harder than we guess today." He also admits that it is "ten times" more effort to understand AI safety risk as developers get "closer and closer" to AGI. [Machine, 2025; Lex Fridman, 2025] CEO Sundar Pichai highlighted the benefits of "recursive self-improvement" for dramatically accelerating creation. [Lex Fridman, 2025] Pichai also highlighted the promising prospect of AI while refining risk language to digital divide, workforce reshaping, and the need for governments to set "rules of the road." [Google, 2026] **Recap from Winter 2025 AI Safety Index** **Quotes in the media from leadership:** **Demis Hassabis (CEO)** - "We must take the risks of AI as seriously as other major global challenges, like climate change [...] It took the international community too long to coordinate an effective global response [..]. We can't afford the same delay with AI" [Guardian, 2024]. - "Artificial intelligence is a dual-use technology like nuclear energy: it can be used for good, but it could also be terribly destructive" [Time, 2025]. - Demis shares that he thinks AGI is only a "handful of years away" and that he is very worried about deception, calling it "incredibly dangerous", and speaks about encouraging the Security institutes to investigate them [Youtube, 2025]. Other examples: [CNN, 2025] [CBS, 2025] Shane Legg (Chief AGI Scientist) communicates a similar stance, and he recently stated AI is a very powerful technology, and it can and should be regulated." [Axios,2025]. In contrast, at the 2025 AI Action Summit in Paris, Google's CEO Sundar Pichai stated that "The biggest risk could be missing out." [Observer, 2024] **CAIS statement on extinction risk signed by Demis Hassabis (CEO), Shane Legg (Co-Founder), Lila Ibrahim (COO).**

Table continues on next page

Meta	*Muse Spark*	**Company and leadership communications on AI risk have historically been muted at Meta but there has been increasing emphasis on catastrophic risks following the 2025 leadership reshuffle.** **Update** The company's updated Advanced AI Scaling Framework (v.2) has increasingly addressed the catastrophic risks brought by AI, expanding risk coverage as well as deepening safety commitments. Prior to joining Meta, Alexandr Wang, the Chief AI Officer who now leads Meta Superintelligence Labs (MSL), argued that "superintelligence is inescapably a matter of national security," recognizing that these systems can "also be turned to destructive ends, enabling rogue actors to engineer bioweapons and hack critical infrastructure."[Hendrycks, Schmidt and Wang, 2025] After he joins Meta, when discussing responsibilities of building personal superintelligence, he argued that "Given how intimately your personal A.I. will know you, people aren't going to hire us for the job if we're not doing it responsibly." [Observer, 2026] On superintelligence, Alexandr emphasizes that it is important to "take superintelligence seriously, and then start to rebuild all of your other assumptions around that core premise." [OfficeChai, 2026] **Recap from Winter 2025 AI Safety Index** Mark Zuckerberg and [former] Chief AI Scientist Yann LeCun express the strongest counter narrative to AI existential risk concerns among major companies [Interesting Engineering, 2025]. Meta's president of global affairs expresses a similar position [Politico, 2024], comparing the discussion and framing the topic as a "moral panic" [Independent, 2024]. Zuckerberg is concerned about power concentration: "But I stay up at night worrying more about an untrustworthy actor having the super strong AI, whether it's an adversarial government or an untrustworthy company or whatever." He shares that:" Bioweapons are one of the areas where the people who are most worried about this stuff are focused, and I think it makes a lot of sense." He expresses less urgency on existential risk addressing deception as "longer-term theoretical risks", and saying ".. we focus more on the types of risks that we see today .." [Dwarkesch Podcast, 2024]. **The recap has removed Yann Lecun's statement as he has left the company in November 2025. [Lecun, 2025]**
xAI	*Grok 4.1*	**Corporate communication itself does not publicly share information about extreme risks. CEO Musk has a track-record of raising concerns.** **Update** Elon Musk argued that "AI and robots will replace all jobs," making working entirely optional. [X, 2025] In the courtroom of the Musk v. Altman trial, he warns the jurors that AI "could kill us all." [MIT Technology Review, 2026] **Recap from Winter 2025 AI Safety Index** Elon Musk argued that "Long-term, AI's gonna be in charge, to be totally frank, not humans. If artificial intelligence vastly exceeds the sum of human intelligence, it is difficult to imagine any humans would actually be in charge" [Pravda, 2025] In 2014, Musk called AI humanity's "biggest existential threat.", calling for regulatory oversight [Live Science, 2014] In September 2023, he told senators "'there's some chance – above zero – that AI will kill us all." [NBC, 2023]. At the 2024 Saudi summit, he estimated a "10-20% chance AI goes bad."[Fortune, 2025] **CAIS statement on AI Risk signed by: Igor Babuschkin (co-founder), Tony Wu (co-founder)**
DeepSeek	*V3.2*	**Researchers and policy teams at the companies are engaging a little more with the topic of existential risks, signaling a more public engagement of the company in the field.** **Recap from Winter 2025 AI Safety Index** DeepSeek researcher Chen Deli struck a conspicuously pessimistic note about the future of AI at a major state-backed tech conference on Friday, warning about its potentially "dangerous" impacts on society and the job market. Chen said he was optimistic about the tech itself but pessimistic about its overall impact on society: "Humans will be completely freed from work in the end, which might sound good but will actually shake society to its core." In September, 2025, DeepSeek's head of AI governance spoke at an open-source conference about ethical guardrails. [SCMP, 2025] The company and its leadership do not discuss extreme risks from AI. CEO Liang Wenfeng keeps a very low profile and rarely speaks in public. Beijing instructed DeepSeek "not to engage with the media without approval." [Reuters, 2025].
Z.ai	*GLM-5*	**Z.ai Corporate communications don't speak about the potential for extreme risks. Leadership has been more actively engaging with the subject.** **Recap from Summer 2025 AI Safety Index** In October 2025, Z.ai's CEO Peng Zhang signed the FLI's superintelligence statement, calling for a prohibition on the development of superintelligence, not lifted before there is broad scientific consensus that it will be done safely and controllably, and strong public buy-in. While corporate communication rarely discusses catastrophic and existential risks, the company's Chief Scientist Tang Jie and its CEO have acknowledged the need to get prepared for existential risks and align super intelligent systems.
Alibaba Cloud	*Qwen3.5*	**Corporate communications and the company's leadership rarely engage with the subject publicly.**
Mistral	*Large 3*	**Corporate communications rarely mention existential risks of AI, and the company's leadership dismisses the notions as well.** CEO of Mistral AI Arthur Mensch says warnings about extreme risks of artificial intelligence are often "distraction tactics," instead arguing that the biggest risk in the near future is that of massive influence on how people think and how they vote. He believes that exaggerated fears about existential risks from AI are being used by some large companies to lobby for regulations that would entrench their own market power and lock out smaller competitors. [Le Monde, 2026] Mensch dismisses the AGI pursuit as a "pseudo-religious endeavor" — "The whole AGI rhetoric is about creating God. I don't believe in God. I'm a strong atheist. So I don't believe in AGI." [Business Insider, 2024] In an earlier interview with Elad Gil, he argued that the discussion around existential risk is "ill defined and has little scientific evidence," including the claim that LLMs are able to generate bioweapons. [Elad Blog, 2024]

Public Policy

Indicator

Policy Engagement on AI Safety Regulations

Definition

This indicator tracks a company's involvement in proactively shaping or responding to laws and regulations concerning AI safety. Evidence includes public statements, consultation submissions, testimony, and official responses, participation in trade associations or coalitions that lobby on safety-related issues, as well as active participation in drafting relevant regulations and standards.

Why it matters

Leading AI developers have unique technical expertise and credibility to advise governments on charting a responsible path for this transformative technology. Tracking patterns in companies' engagements on specific regulations can indicate which firms take a proactive stance on raising the bar for sensible protections.

Anthropic	**Update** Anthropic announced a $20 million donation to Public First Action, a group supporting AI guardrails. The group plans to back 30-50 candidates from both parties in state and federal races. [CNBC, 2026] **Illinois SB 315** Anthropic has publicly endorsed the bill. [NBC News, 2026] **Recap from Winter 2025 AI Safety Index** **California SB 53** SB 53 provides "a blueprint for evidence-generating transparency measures" for governing frontier AI systems. [Carnegie Endowment, 2025] Anthropic publicly endorsed SB 53, calling it a "trust-but-verify" approach that strengthens accountability for frontier AI systems and sets a strong baseline for transparency. The company emphasized that while it still prefers a federal framework, California's action is necessary given the rapid development of advanced models [Anthropic, 2025; TechCrunch, 2025]	**Preemption of state-level AI legislation** In its endorsement announcement for California SB 53, it stated that "frontier AI safety is best addressed at the federal level instead of a patchwork of state regulations," deviating from its previous stance on state-oriented AI safety approach. **EU** *EU AI Act* N/A **US Legislations** *California SB 1047* Anthropic raised initial concerns about key provisions, but the CEO later expressed cautious support, acknowledging that the benefits of the bill likely outweigh its costs. It also actively shapes the final version of the legislation. *New York Raise Act* N/A *Preemption of state-level AI legislation* In 2025, Anthropic opposed federal efforts to preempt state-level AI laws. CEO Dario Amodei argued that states should retain authority to set transparency and safety standards, warning that federal preemption could weaken oversight.
OpenAI	**Update** OpenAI's co-founder Greg Brockman co-founded *Leading the Future* with Andreessen Horowitz, which has raised $125M+ since mid-2025. The group explicitly aims to combat "burdensome regulation" including pre-release safety testing and liability requirements. Leading the Future has already spent almost $16 million (as of May 2026) through two subsidiaries to back candidates from both parties who support a national regulatory framework for AI. Chris Lehane, OpenAI's chief of global affairs, clarifies that it does not directly fund any super PACs. [Wired, 2026; WSJ, 2026; New York University, 2026] **Illinois SB 315** OpenAI endorsed Illinois' frontier AI safety bill, SB 315, arguing that it represents something bigger than a single state effort, while still recognizing the importance of having a "common baseline" amongst states. [OpenAI, 2026] **Recap from Winter 2025 AI Safety Index** **California SB 53** No public stance **EU** *EU AI Act* In 2023, OpenAI lobbied EU officials to weaken parts of the AI Act, arguing that foundation models such as GPT-4 should not face strict obligations unless adapted for specific uses. [TIME, 2023]	**US Legislations** *California SB 1047* In 2024, OpenAI opposed California's SB 1047, arguing that its safety requirements—such as third-party evaluations and incident reporting—would hinder innovation and disadvantage U.S. firms. [Bloomberg, 2024] *New York Raise Act* N/A *Preemption of state-level AI legislation* In 2025, OpenAI supported federal preemption of state-level AI laws, arguing that a unified national framework would better promote innovation and avoid regulatory fragmentation. In OpenAI's letter to Governor Newsom on harmonized regulation, the company urges California to "harmonize" with federal and global frameworks instead of layering its own additional requirements. OpenAI argued that "a patchwork of state rules… could slow innovation without improving safety," urging California instead to align with "federal and global safety guidelines" to "avoid duplication and inconsistencies between state requirements and the safety frameworks already being advanced by the US government and our democratic allies." [OpenAI, 2025]

Table continues on next page

Google DeepMind	**Update** **Illinois SB 315** The Computer & Communications Industry Association (CCIA), of which Google is a member, has opposed the bill. [CCIA, 2026] **Recap from Winter 2025 AI Safety Index** **California SB 53** Industry group TechNet that represents Google opposed SB 53, arguing that the bill's scope is too broad and that the disclosure and reporting requirements could expose trade secrets or magnify security vulnerabilities. [Citizen Portal, 2025] [San Francisco Standard, 2025] **EU** *EU AI Act* Google DeepMind opposed classifying general-purpose and foundational models as "high-risk," arguing this would stifle innovation and that regulation should target downstream applications.	**US Legislations** *California SB 1047* Google DeepMind opposed California's SB 1047, arguing that its safety rules would burden developers and stifle innovation, and that state oversight can fragment regulation. *New York Raise Act* Industry groups with ties to Google opposed the RAISE Act, arguing that the legislation could conflict with federal policy and impose overly broad restrictions on AI development. *Preemption of state-level AI legislation* In its response to the U.S. AI Action Plan in 2025, it called for federal leadership over issues like copyright, export controls, and development standards, warning that state-level rules could hinder innovation.
Meta	**Update** **Illinois SB 315** The Computer & Communications Industry Association (CCIA), of which Meta is a member, has opposed the bill. [CCIA, 2026] **Recap from Winter 2025 AI Safety Index** **California SB 53** No public stance **EU** *EU AI Act* Between 2022 and 2023, Meta lobbied EU institutions to limit safety rules in the AI Act, opposing strict obligations for general-purpose models and seeking exemptions for open-source systems.	**US Legislations** *California SB 1047* In 2024, Meta lobbied against California's SB 1047, arguing that its AI safety requirements—especially pre-deployment risk assessments and licensing—were overly broad and could hinder innovation. *New York Raise Act* In 2025, Meta opposed RAISE Act through multiple affiliated groups, including Tech:NYC, the AI Alliance, and the Computer & Communications Industry Association. *Preemption of state-level AI legislation* In 2025, Meta advocated for federal preemption of state-level AI regulations, warning that fragmented laws could create compliance challenges and hinder innovation across jurisdictions.
xAI	**Update** **Illinois SB 315** No public stance. **Recap from Winter 2025 AI Safety Index** **California SB 53** No public stance **EU** *EU AI Act* No public stance	**US Legislations** In 2024, xAI CEO Elon Musk publicly supported the bill in an X post. *New York Raise Act* No public stance *Preemption of state-level AI legislation* No public stance.
DeepSeek	DeepSeek is among the entities that have drafted the **Cybersecurity technology—Basic security requirements for generative artificial intelligence service**, which is a voluntary national standard that focuses on safety requirements including corpus safety, model safety, and safety assessment, although it doesn't mention frontier AI risks. [National Service Platform for Standards Information]	
Z.ai		
Alibaba Cloud	Alibaba is among the entities that have drafted the **Cybersecurity technology—Labeling method for content generated by artificial intelligence**, which is the only binding national standard that requires generative AI services to label AI-generated content both in implicit and explicit ways. [National Public Service Platform for Standards Information]	
Mistral	**EU** The EU AI Act is far from ideal, but regulation is currently in a "workable state" and isn't the biggest problem for Europe right now, according to Arthur Mensch, the CEO of Mistral. [MLex, 2025]	

TO BE COMPLETED BY PANELLISTS

Grading Sheet: Information Sharing and Public Messaging

Please pick a grade for each firm. You may use the full letter-grade scale with +/− modifiers as appropriate. You can add brief justifications to your grades.

	Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud	Mistral
Grades									
Grade comments (Justifications, opportunities for improvements, etc.)									

Grading Scales

Grading scales are provided to support consistency between reviewers.

- **A** Provides detailed, verifiable disclosures on model safety and governance; fully cooperates with external evaluations; publicly and legislatively advocates for stronger safety and accountability standards.
- **B** Shares clear information on key safety and governance aspects; engages with external processes; publicly supports most safety initiatives while maintaining some self-interest.
- **C** Offers limited or curated safety and governance information; selectively participates in external efforts; adopts mixed or neutral positions on safety regulation.
- **D** Rarely discloses meaningful information; limited or inconsistent cooperation; messaging downplays risks or discourages stronger oversight.
- **F** Withholds or distorts safety information; no credible cooperation; messaging actively undermines safety regulation or misleads the public on risk.

Domain comments

Optional: Share observations that apply across companies, including general recommendations, notes on how you weighted indicators, or feedback on FLI's methodology.

Domain comments	

Appendix B: Company Survey

Introduction

Thank you for participating in the **FLI AI Safety Index 2026 Survey**. This survey is designed to allow your company to provide additional information about specific practices and policies for managing risks from advanced AI systems. The independent experts on the review panel will consider the information you provide here when evaluating your company's safety efforts.

Survey instructions

The survey contains a total of **34 questions**, which predominantly follow a **multiple-choice format**. Where options are provided, select the one that best fits your current practices. Some questions allow a brief explanation or ask for details (especially if you answered "Other" or an open-ended part) – please be concise and factual in those responses. You are welcome to provide **URLs or document references** for any publicly available policies or reports that support your answers. It is not necessary to answer all questions within the survey. You can skip specific questions when answering would be difficult/inconvenient.

You have received a personalized link which you can share with colleagues to collaborate on the survey. You do not need to fill out the survey in a single sitting. Progress will be saved whenever you navigate between sections.

Confidentiality

Please do not share confidential information. We plan to publish all survey responses in full after the grading process is completed.

We appreciate your time and effort in providing thorough answers.

Whistleblowing policies (16 Questions)

If your company has region-specific whistleblowing (WB) policies instead of a single global WB policy, please answer all questions in this survey with regard to the policy that applies to the majority of your frontier AI-focused management, research, and engineering employees. Unless a question specifically asks about other stakeholders, please answer based on protections available to current full-time employees. You may explain variations for different stakeholder groups in the final question.

You can use the text-box at the end of this section to provide clarifications and/or link to relevant publicly available documents.

Definition of terms:

Whistleblowing Function:

The organizational structure, personnel, processes, and resources established to receive, assess, investigate, and respond to whistleblowing reports. This includes the designated individuals or teams responsible for writing and acting according to the whistleblowing policy, managing the whistleblowing process, any technological systems used to facilitate reporting, and the mechanisms for investigating and addressing reported concerns.

Whistleblowing Policy:

The formal, documented set of rules, procedures, and guidelines that govern how an organization handles whistleblowing. This policy outlines what concerns can be reported ("material scope"), who can report them ("covered persons"), how reports should be made and to whom, how they will be handled, and what protections are available to whistleblowers who follow this policy. It serves as the official framework that defines the organization's approach to whistleblowing.

Covered persons:

Individuals who are explicitly protected when making good-faith reports under the whistleblowing policy. The range of covered persons may vary by organization and jurisdiction.

Material scope:

The range of issues, concerns, violations, or misconduct that can legitimately be reported through the whistleblowing channels and will be considered for investigation. In this context, this may include legal violations, ethical breaches, safety concerns, alignment issues, misrepresentations of capabilities, or other matters related to responsible AI development and deployment that the organization has defined as reportable concerns.

Question Title	Available options	OpenAI	Z.ai	Anthropic	Google Deepmind	Meta
Does your company have a WB policy & function covering frontier AI-focused staff? *Is this policy publicly accessible without login credentials? - Selected Choice*	Prefer not to answer (skips whistleblowing section) ▪ No WB policy & function - (skips whistleblowing section) ▪ Non-public policy exists - Please briefly explain your rationale for keeping it private:	Public WB policy - Please provide URL here: https://cdn.openai.com/policies/openai-raising-concerns-policy.pdf	No changes since last year (skips whistleblowing section) - please specify anything you'd like to clarify since last year.	Public WB policy - Please provide URL here: We posted an update to our RSP Non-compliance Reporting and Anti-Retaliation Policy, which we released internally in February 2026. It expands reporting channels, introduces a pathway for employees to make informal inquiries about potential RSP violations, and aligns with RSP Version 3.0. The policy document can be found at the link below. https://www-cdn.anthropic.com/b7a5629e40b391b2adf-b4cc8c0888ac9d6bfddf6/RSP%20Non-compliance%20Reporting%20and%20Anti-Retaliation%20Policy.pdf In addition to our RSP Noncompliance Reporting and Anti-Retaliation Policy, we maintain an internal whistleblowing policy that broadly covers potential violations of law, company policies, and our standards.	No changes since last year (skips whistleblowing section) - please specify anything you'd like to clarify since last year.	Non-public policy exists - Please briefly explain your rationale for keeping it private: Meta maintains a comprehensive Whistleblower and Complaint Policy, available to all Meta Personnel and summarized externally in our Code of Conduct available here: https://www.meta.com/people-practices/code-of-conduct/?srsltid=AfmBOor3aZo81GVFszSRugRQvwDkqqAybKArhrjiM34CFiKDGejgZTZh

Table continues on next page

Question Title	Available options	OpenAI	Z.ai	Anthropic	Google Deepmind	Meta
Who is formally designated with primary responsibility for overseeing the whistleblowing function and ensuring reports are properly addressed? - Selected Choice	▪ Board/Audit Committee ▪ Executive management ▪ Compliance/Legal department ▪ HR department ▪ Other (Please also specify whom this role reports to):	Other (Please also specify whom this role reports to): The Board, Compliance/ Legal department and the HR department each have roles as described in our published policy.		Compliance/Legal department Meta maintains a comprehensive whistleblower and complaint policy, and is developing further protocols to report any instances of non-compliance with our Advanced AI Scaling Framework, and any specific and substantial danger to the public health or safety arising from catastrophic risk. Under this protocol, employees will be able to confidentially, and, if they choose, anonymously issue reports through internal channels, and all reports will be ultimately submitted to the internal governance function, the Chief AI Officer, and the Director of Alignment and Risk.		Other (Please also specify whom this role reports to):
Which statement best describes the investigative independence of your whistleblowing function?	▪ The whistleblowing function requires approval from management before initiating investigations based on whistleblower reports. ▪ The whistleblowing function can independently initiate and conduct investigations based on whistleblower reports, including those involving senior management. ▪ The whistleblowing function can independently initiate and conduct investigations based on whistleblower reports, including those involving senior management, AND has the authority to engage external expertise without approval.	The whistleblowing function can independently initiate and conduct investigations based on whistleblower reports, including those involving senior management, AND has the authority to engage external expertise without approval.		The whistleblowing function can independently initiate and conduct investigations based on whistleblower reports, including those involving senior management, AND has the authority to engage external expertise without approval.		The whistleblowing function can independently initiate and conduct investigations based on whistleblower reports, including those involving senior management, AND has the authority to engage external expertise without approval.
Which of the following concerns are explicitly covered by your whistleblowing policy? (Select all that apply)	▪ Violations of applicable laws and regulations ▪ Violations of the company's public AI safety framework (e.g., Anthropic's Responsible Scaling Policy) ▪ Credible safety concerns that may not violate specific policies including loss-of-control scenarios ▪ Pressure to compromise safety standards or suppress safety concerns ▪ Misleading communications about AI capabilities to external parties (such as regulators, the public, or evaluators) or discrepancies between public claims and internal practices ▪ None of the above	Violations of applicable laws and regulations,Violations of the company's public AI safety framework (e.g., Anthropic's Responsible Scaling Policy),Credible safety concerns that may not violate specific policies including loss-of-control scenarios,Pressure to compromise safety standards or suppress safety concerns		Violations of applicable laws and regulations,Violations of the company's public AI safety framework (e.g., Anthropic's Responsible Scaling Policy),Credible safety concerns that may not violate specific policies including loss-of-control scenarios,Pressure to compromise safety standards or suppress safety concerns,Misleading communications about AI capabilities to external parties (such as regulators, the public, or evaluators) or discrepancies between public claims and internal practices		Violations of applicable laws and regulations,Violations of the company's public AI safety framework (e.g., Anthropic's Responsible Scaling Policy),Credible safety concerns that may not violate specific policies including loss-of-control scenarios,-Pressure to compromise safety standards or suppress safety concerns,Misleading communications about AI capabilities to external parties (such as regulators, the public, or evaluators) or discrepancies between public claims and internal practices
Does your whistleblowing policy explicitly protect individuals who report concerns in 'good faith' or with 'reasonable cause to believe,' rather than requiring certainty that violations occurred?	▪ Yes ▪ No	Yes		Yes		Yes

Table continues on next page

Question Title	Available options	OpenAI	Z.ai	Anthropic	Google Deepmind	Meta
Which of the following persons are protected from retaliation under your whistleblowing policy? (Select all that apply)	• Current employees • Former employees • Contractors and self-employed workers • AI research collaborators and academic partners • Individuals who assist whistleblowers • Suppliers and vendors with access to company systems	Current employees,Contractors and self-employed workers,AI research collaborators and academic partners,Individuals who assist whistleblowers,Suppliers and vendors with access to company systems		Current employees,Former employees,Contractors and self-employed workers,AI research collaborators and academic partners,Individuals who assist whistleblowers,Suppliers and vendors with access to company systems		Current employees,Former employees,Contractors and self-employed workers,AI research collaborators and academic partners,Individuals who assist whistleblowers,-Suppliers and vendors with access to company systems
To which of the following individuals or entities can whistleblowers submit reports according to your policy? (Select all that apply) - Selected Choice	• Board member or board committee • Dedicated Ethics/Whistleblowing Officer • Ombudsperson • Chief Compliance or Risk Officer • General Counsel/Legal Department • Human Resources department • External/independent third party • Direct disclosure to a statutory or supervisory authority • Other (please briefly specify):	Board member or board committee,Chief Compliance or Risk Officer,General Counsel/Legal Department,Human Resources department,External/independent third party,Direct disclosure to a statutory or supervisory authority		Board member or board committee,Dedicated Ethics/Whistleblowing Officer,Chief Compliance or Risk Officer,General Counsel/Legal Department,Human Resources department,External/independent third party,Direct disclosure to a statutory or supervisory authority		
For former employees and contractors, indicate any policy limitations compared with current employees. (Select all limitations that apply)	• Limited Reporting Channels • Limited Reportable Issues • Limited Retaliation Protection • No Limitations • For each, specify whether the limitation applies to: • Former employees • Contractors	Some channels, such as speaking to your current HR representative, are inherently available only to current employees		No limitations: Former employees No limitations: Contractors		
Which of the following best describes the anonymity and confidentiality provisions in your whistleblowing policy? (Select the one that fits best)	• Our policy does not provide for anonymous reporting • Our policy allows anonymous reporting but does not specify technical measures to protect reporter identity • Our policy allows anonymous reporting with specific technical measures in place to protect reporter identity (e.g., anonymous hotline, encrypted system) • Our policy allows anonymous reporting with technical protections AND includes confidentiality commitments for non-anonymous reports	Our policy allows anonymous reporting with technical protections AND includes confidentiality commitments for non-anonymous reports		Our policy allows anonymous reporting with technical protections AND includes confidentiality commitments for non-anonymous reports		

Table continues on next page

Question Title	Available options	OpenAI	Z.ai	Anthropic	Google Deepmind	Meta
Does your whistleblowing policy explicitly protect employees disclosing to external parties (e.g., regulators, accredited journalists, civil-society groups) when internal channels are unavailable, conflicted, or fail to resolve a serious concern within stated timelines? (Select one) *Possible Conditions:* ▪ *Imminent risk of serious harm* ▪ *Management or board implicated* ▪ *Reasonable fear of retaliation* ▪ *Internal investigation deadlines missed* ▪ *Unconditional reporting to a competent regulatory authority* ▪ *After internal reporting has been attempted*	▪ No – external disclosure is not explicitly protected or is discouraged (skips follow-up question) ▪ Limited – protected only under specific conditions (choose below) ▪ Full – broadly protected under all listed conditions above (skips follow-up question)	Full – broadly protected under all listed conditions above (skips follow-up question) Note: Our policy specifically protects disclosures to any "national, federal, state or local agency charged with the enforcement of any laws or regulations."		Full – broadly protected under all listed conditions above (skips follow-up question)		
If "Limited", under which circumstances is external disclosure protected? - Selected Choice	▪ Imminent risk of serious harm ▪ Management or board implicated ▪ Reasonable fear of retaliation ▪ Internal investigation deadlines missed ▪ Unconditional reporting to a competent regulatory authority ▪ After internal reporting has been attempted ▪ Other (specify):					
Which mechanisms ensure that your whistleblowing function has access to adequate (technical) expertise to investigate reports? (Select all that apply) - Selected Choice	▪ Dedicated AI experts within the whistleblowing function itself ▪ Authority to consult internal AI experts under confidentiality safeguards, including procedures that shield case details where necessary ▪ Standing agreements with external independent AI ethics/safety consultants ▪ Budget authority to engage external AI experts without requiring management approval ▪ None of the above ▪ Other (please specify):	Authority to consult internal AI experts under confidentiality safeguards, including procedures that shield case details where necessary		Dedicated AI experts within the whistleblowing function itself,Authority to consult internal AI experts under confidentiality safeguards, including procedures that shield case details where necessary		Authority to consult internal AI experts under confidentiality safeguards, including procedures that shield case details where necessary
Investigation timelines and escalation rights: Which best describes your policy's commitments? (Select one)	▪ None – no specific timelines for acknowledgment, updates, or resolution ▪ Basic – acknowledge receipt ≤ 7 days only ▪ Standard – acknowledge ≤ 7 days and provide updates ≤ 30 days ▪ Full – acknowledge ≤ 7 days, updates ≤ 30 days, final outcome ≤ 90 days ▪ Full + internal escalation – all Full timeframes plus whistleblowers may escalate to board/leadership if deadlines are missed ▪ Full + comprehensive escalation – all Full timeframes plus whistleblowers may escalate both internally AND to regulators/external parties if deadlines are missed	Standard – acknowledge ≤ 7 days and provide updates ≤ 30 days				

Table continues on next page

Question Title	Available options	OpenAI	Z.ai	Anthropic	Google Deepmind	Meta
Which specific forms of retaliation are explicitly prohibited in your policy? (Check all that apply)	▪ Termination/Dismissal ▪ Demotion, or negative performance reviews ▪ Reduction in compensation or benefits ▪ Exclusion from meetings or information ▪ Harassment or creating a hostile work environment ▪ Blacklisting within the industry ▪ Legal action against the whistleblower ▪ None of the above	Our policy forbids retaliation. Notwithstanding the way this question is worded, it is well established under relevant law that retaliation can include termination or dismissal, demotion or negative performance reviews, or reduction in compensation or benefits. These are all covered under our policy's prohibition of retaliation. Our policy also expressly addresses harassment.		Termination/Dismissal,Demotion, or negative performance reviews,Reduction in compensation or benefits,Exclusion from meetings or information,Harassment or creating a hostile work environment,Blacklisting within the industry,Legal action against the whistleblower		
Do any employment-, separation-, or settlement-related agreements used by your company contain non-disparagement or confidentiality clauses that could deter current or former employees from disclosing AI safety or risk-related concerns? (Select one)	▪ No - we do not include such restrictions in our agreements ▪ Yes, but clauses only limit public disclosure; internal or regulator disclosures are explicitly unrestricted. ▪ Yes, but not enforced – clauses exist, but the company has a written policy never to enforce (or threaten to enforce) them against AI safety or risk-related disclosures (no withholding of pay/equity and no legal action). ▪ Yes, enforced - our standard confidentiality and non-disparagement provisions may restrict raising AI safety or risk-related concerns	Yes, but clauses only limit public disclosure; internal or regulator disclosures are explicitly unrestricted. We have confidentiality clauses that could impact some forms of public disclosure, but these have carveouts for internal or regulator disclosures. We do not have non-disparagement clauses in any such agreements, except in specific cases where an employee or former employee has entered a mutual non-disparagement agreement with the company.		No - we do not include such restrictions in our agreements		
Which anti-retaliation provisions are explicitly detailed in your whistleblowing policy? (Select all that apply)	▪ Defined disciplinary consequences for individuals who retaliate against whistleblowers (e.g., termination, demotion, or other concrete penalties - not just general statements prohibiting retaliation) ▪ Documented investigation procedure for retaliation claims (including designated investigators, timelines, evidence standards, and appeal rights) ▪ Concrete remedial measures for whistleblowers who experience retaliation (e.g., compensation, reinstatement, transfer options, or other specific remedies - not just general commitments to address retaliation) ▪ None of the above are specifically detailed	Defined disciplinary consequences for individuals who retaliate against whistleblowers (e.g., termination, demotion, or other concrete penalties - not just general statements prohibiting retaliation)		Defined disciplinary consequences for individuals who retaliate against whistleblowers (e.g., termination, demotion, or other concrete penalties - not just general statements prohibiting retaliation)		

External Pre-Deployment Safety Testing (6 Questions)

Please answer the following questions about external pre-deployment safety testing with regards to the release of your currently most capable publicly deployed AI model.

Frontier models:

- Anthropic - Claude Sonnet 4.5
- DeepSeek - R1
- Google Deepmind - Gemini 2.5 Pro
- Meta - Llama 4 Maverick
- OpenAI - GPT-5
- xAI - Grok-4
- Z.ai - GLM-4.6
- Alibaba Cloud - Qwen 3 Max

You can use the text-box at the bottom of the page to provide clarifications and/or link to relevant publicly available documents.

Question Title	Available options	OpenAI	Z.ai	Anthropic	Google Deepmind	Meta
Did your organisation commission one or more independent (no financial/ governance ties to your company) organisations to test this model for the dangerous capabilities or propensities you prioritized (in safety framework if available) before public release? - Selected Choice	▪ No – no such external pre-deployment testing was commissioned (skip to next section) ▪ Yes – external testing was commissioned. Please list the organization(s) that performed relevant tests on the specified model and briefly indicate the broad risk domain(s) covered e.g., "UK AISI: cyber-offense, bio-risk" (opens follow-up questions below):	Yes – external testing was commissioned. Please list the organization(s) that performed relevant tests on the specified model and briefly indicate the broad risk domain(s) covered e.g., "UK AISI: cyber-offense, bio-risk (opens follow-up questions below): We've worked with the US CAISI and the UK AI Security Institute, independent third party labs such as METR, Apollo Research, SecureBio and Irregular to add an additional layer of validation for key risks. Where possible and relevant, we report on their findings in our system cards, and empower external testers to publish their own findings. Most recently, in the GPT-5.5 System Card (https://deploymentsafety.openai.com/gpt-5-5) we detail external capability evaluations by Apollo for sandbagging, by SecureBio and US CAISI for biological capabilities, by Irregular, US CAISI and UK AISI for cyber capabilities, and by external redteamers for our cyber safeguards. Third party assessors were provided OpenAI GPT-5.5 early checkpoints, as well as the final launch candidate models to conduct their assessments across Cyber, Bio, and Sandbagging . As part of our ongoing efforts to consult with external experts, OpenAI granted early access to these versions of GPT-5.5 to both CAISI and UK AISI, both who conducted evaluations of the model's cyber and biological and chemical capabilities, as well as safeguards. As part of a longer-term collaboration, UK AISI was also provided access to prototype versions of our safeguards and information sources that are not publicly available – such as our monitor system design, biological content policy, and chains of thoughts of our monitor models. This allowed them to perform more rigorous stress testing and identify potential vulnerabilities more easily. Gray Swan and FAR.AI conducted universal jailbreak red teaming for key frontier risks. Apollo Research evaluated in-context scheming and strategic deception. Irregular evaluated the model's cybersecurity related capabilities, and SecureBio evaluated the models' biological capabilities. METR chose not to directly evaluate GPT-5.5 due to due capacity constraints and prioritization in light of a broader pilot collaboration already underway with OpenAI, which involves a longer term assessment for AI R&D and loss of control risks on a regular cadence, and they were provided early access with GPT-5.5.	No – no such external pre-deployment testing was commissioned (skip to next section)	Yes – external testing was commissioned. Please list the organization(s) that performed relevant tests on the specified model and briefly indicate the broad risk domain(s) covered e.g., "UK AISI: cyber-offense, bio-risk (opens follow-up questions below): UK AISI - cyber Please see our system cards (library, Claude Opus 4.7) and transparency hub for information on our external testing https://docs.claude.com/en/resources/Overview, https://cdn.sanity.io/files/4zrzovbb/website/037f06850df7fbe871e206dad004c3db5fd50340.pdf, https://www.anthropic.com/transparency/voluntary-commitments	Yes – external testing was commissioned. Please list the organization(s) that performed relevant tests on the specified model and briefly indicate the broad risk domain(s) covered e.g., "UK AISI: cyber-offense, bio-risk (opens follow-up questions below): Yes, external safety testing was commissioned for 3.1 Pro, including across CBRN, Loss of Control, Cyber, and Harmful Manipulation. We have worked with a diverse group of external experts, including Apollo Research, Dreadnode and Vaultis.	Yes – external testing was commissioned. Please list the organization(s) that performed relevant tests on the specified model and briefly indicate the broad risk domain(s) covered e.g., "UK AISI: cyber-offense, bio-risk (opens follow-up questions below): ScaleAI: Model resistance to harmful requests HandshakeAI: Robustness against jailbreak Apollo Research: Evaluation awareness, deceptive alignment, loss of control Irregular: cybersecurity evaluations Deloitte, Faculty AI, SecureBio, and Frontier Design: Engagements for biodefense and biosecurity, including workflows for threat modeling, experimental design and testing, and the interpretation and validation of evaluation results

Table continues on next page

Question Title	Available options	OpenAI	Z.ai	Anthropic	Google Deepmind	Meta
For each organization listed in Q17, how would you characterize their independence from your company? (Select all that apply across your listed evaluators. If different evaluators had different arrangements, please describe in the text box.) - Selected Choice		One or more evaluators had financial, governance, or other dependencies not described above — please specify, including which organization(s) this applies to: (note - "Other" was automatically selected above because we are using this text box) No financial ties: US CAISI, UK AISI, METR No governance ties: US CAISI, UK AISI, Irregular, Apollo. SecureBio, METR No material commercial dependency: US CAISI, UK AISI, METR We provide API credits to all third party assessors as needed to help facilitate their research and evaluations. We and they are not considering this as a financial incentive given that the credits are granted and principally used for these evaluations. We share more about this, our legal agreements, and other principles for third party testing in our blog: https://openai.com/index/strengthening-safety-with-external-testing/			For each external tester we work with for pre-deployment testing, we undergo a due diligence process to ensure their independence. This covers, among other areas, governance and commercial dependency. GDM does provide payment for conducting evaluations, to ensure they are fairly compensated for their time. We have measures in place to ensure that these payments are in no way conditional on the conclusions they reach.	
What was the highest level of technical access granted to any of the listed external evaluators during pre-deployment testing for the specified release? (Select the highest level that applies)	▪ Standard inference API with normal user-facing filters in place ▪ Inference API with safety filters disabled (no inference-time mitigations) ▪ "Helpful-only" or base model API (no harmlessness fine-tuning and no filters) ▪ Fine-tuning interface without safety gatekeeping ▪ Direct read/write access to internal activations or weights	"Helpful-only" or base model API (no harmlessness fine-tuning and no filters) We also share visible Chain of Thought access for evaluators who require this.			Inference API with safety filters disabled (no inference-time mitigations) External testing partners were provided the model without inference time mitigations relevant to their specific domain. Answers for Gemini 3.1 Pro	"Helpful-only" or base model API (no harmlessness fine-tuning and no filters)
What was the longest period of time that an external evaluator was given continuous access for pre-deployment testing of your model? (Select one)	▪ >5 weeks ▪ >3 weeks ▪ >2 weeks ▪ >1 week ▪ <1 week	>2 weeks			>2 weeks	

Table continues on next page

Question Title	Available options	OpenAI	Z.ai	Anthropic	Google Deepmind	Meta
Which of the following publication arrangements applied to external evaluators' findings? *If different evaluators had different publication terms, please select all that occurred and briefly explain using the text-box.* *(select all that apply)* *- Selected Choice*	▪ Evaluators may publish independently without prior company approval after the model is released. ▪ Evaluators may publish independently after company review/possible redaction. ▪ The company pre-committed to reproduce an independently written report in the model card without redactions. ▪ The company publishes report after review/possible redactions. ▪ The company provided its own summary of the evaluator's key findings. ▪ Findings remain internal ▪ Other: Please briefly explain:	Evaluators may publish independently without prior company approval after the model is released.,Evaluators may publish independently after company review/possible redaction.,The company publishes report after review/possible redactions.,The company provided its own summary of the evaluator's key findings.,Findings remain internal (a) "Evaluators may publish independently without prior company approval after the model is released" :This is true if they run their evaluations independently on the deployed model. Results from the pre-deployment evaluation period are under NDA / require prior approval to protect confidential information. We share more about our NDA in our blog post here: https://openai.com/index/strengthening-safety-with-external-testing/ (b) "Evaluators may publish independently after company review/possible redaction." See above, in cases where the evaluator wishes to publish about the specifics of the pre-deployment period - METR as an example did publish and made a note that they believe that our redactions did not substantively change their conclusions ("We did not make changes to conclusions, takeaways or tone (or any other changes we considered problematic) based on their review."). Other examples of work being independently published include: UK AISI, Secure Bio. (c) "The company publishes report after review/possible redactions." (d) OpenAI publishes excerpts from the report mutually agreed upon or written, with OpenAI having the final say for what content goes in System Cards. (e) "The company provided its own summary of the evaluator's key findings." This is true in some cases, but we also share back any summaries that we plan to publish with the evaluator prior to release to confirm factual accuracy. (f) "Findings remain internal" - Some evaluators prefer that their full findings are not shared publicly, such as some forms of government testing by the US CAISI and UK AISI.		Evaluators may publish independently after company review/ possible redaction.,The company provided its own summary of the evaluator's key findings.	The company provided its own summary of the evaluator's key findings. GDM publishes high level summaries appropriate for the risks being evaluated within the Models Cards / Tech report with GDM having the final say for what content goes in the Model Cards/Tech report. Answers for Gemini 3.1 Pro	
During pre-deployment testing, what best describes the query-rate or volume restrictions applied to external evaluators? *(Select one)*	▪ No limits – evaluators could automate or batch queries with no additional throttling or hard caps. ▪ Elevated but capped – evaluators had higher quotas than the public/enterprise tier but were still subject to explicit caps (e.g., requests-per-minute or daily token limits). ▪ Public-tier caps – evaluators were held to the same rate/volume limits as ordinary paying users. ▪ Lower than Public-tier caps - evaluators had lower quotas than ordinary paying users.	Elevated but capped – evaluators had higher quotas than the public/enterprise tier but were still subject to explicit caps (e.g., requests-per-minute or daily token limits). Query rates can depend on technical feasibility in some cases.			Elevated but capped – evaluators had higher quotas than the public/enterprise tier but were still subject to explicit caps (e.g., requests-per-minute or daily token limits). Query rate is bespoke depending on the testing partner's specific needs and evaluation type. Where required, GDM provided elevated but capped quotas, but this rate often depended on technical feasibility. Answers for Gemini 3.1 Pro	Elevated but capped – evaluators had higher quotas than the public/ enterprise tier but were still subject to explicit caps (e.g., requests-per-minute or daily token limits).
Does your organization log and retain the model interactions of external evaluators during pre-deployment testing? *- Selected Choice*	▪ Yes - Inputs and outputs are logged and retained. ▪ No - Inputs and outputs are neither logged nor retained, protecting evaluator IP. ▪ Other (please describe):	Other (please describe): Zero Data Retention available upon request, if technically feasible during pre-deployment periods (for some new models or products, ZDR is not always possible during pre-deployment testing).			No - Inputs and outputs are neither logged nor retained, protecting evaluator IP. No - Inputs and outputs are not logged during pre-deployment testing by external evaluators. However, where agreed, external evaluators share prompts and model responses for the purpose of assessment and mitigation of risks. Answers for Gemini 3.1 Pro	

Internal Deployments (3 Questions)

Deployment levels:

1. Broad deployment: Many teams within the company have access for normal use.
2. Development access: Access limited to specific teams or projects that are actively testing the model or developing it further.

Question Title	Available options	OpenAI	Z.ai	Anthropic	Google Deepmind	Meta
If you specified external pre-deployment safety evaluations in the previous section, were these performed before or after broad internal deployment? (Select one)	▪ Before - External safety tests were completed before broad internal deployment. ▪ Partial - All external evaluations on situational awareness, scheming, and cyber-offense were conducted before broad internal deployment. ▪ After - External safety tests were completed after broad internal deployment. ▪ Other (please explain briefly):	After - External safety tests were completed after broad internal deployment.				Before - External safety tests were completed before broad internal deployment.
What level of safety testing does your company require for broad internal deployment of frontier AI models? (Select one)	▪ No formal risk management requirements for internal deployments ▪ Formalized risk management for internal deployments with less stringent requirements than external deployment framework for the following risks/ capabilities: situational awareness, scheming, AI R&D, cyber-offense. ▪ Formalized risk management for internal deployments with the same requirements as external deployment framework for the following risks/capabilities: situational awareness, scheming, cyber-offense. ▪ Company requires the same risk management effort for internal and external deployments. ▪ Other (Please briefly describe):	Other (Please briefly describe): We believe that as models reach High cybersecurity capability, internal deployment itself becomes a meaningful surface to consider – not because of misuse, but because high cyber capability can remove a key bottleneck to certain internal deployment risks materializing. For example, in conjunction with additional capabilities such as long range autonomy, a model with the propensity to self-exfiltrate or sabotage internal research could plausibly succeed at these attempts. This risk makes it important to mature our internal deployment posture ahead of more capable systems. Accordingly, we are accelerating work across evaluations, safeguards, and operating procedures. We recently publicly described how we are monitoring internal coding agents for misalignment (https://openai.com/index/how-we-monitor-internal-coding-agents-misalignment/), with a low-latency internal monitoring system that uses a frontier-class reasoning model at maximum reasoning effort. It reviews the agent's interactions and alerts us to actions that may be inconsistent with a user's intent, or that may violate our own internal security or compliance policies.The monitoring system privately runs across all standard internal coding agent deployments, viewing the full conversation history, including all chains of thought, all user, developer, and assistant messages, along with tool calls and outputs.	Formalized risk management for internal deployments with less stringent requirements than external deployment framework for the following risks/capabilities: situational awareness, scheming, AI R&D, cyber-offense.			
Does your company require any of the following safeguards for broad internal deployments of frontier AI models? *(Select all that apply)*	▪ Inference time safety mitigations for misuse risks (including cyber & bio risks) ▪ Restricting access to helpful-only models and only granting time-bound access to staff that apply with a legitimate research need ▪ Logging all inputs and outputs from internal use and retaining them for at least 30 days ▪ Not currently logging, but introduced an *official, written* plan to start doing so after models reach a specified capability threshold ▪ Analyzing all internal model interactions for abnormal activity, including harmful use or unexpected attempts by AI systems to take real-world actions ▪ Live monitoring and automated editing/resampling of suspicious outputs ▪ None of the above ▪ Other (please describe briefly):"	See answer to Q24, above.	Inference time safety mitigations for misuse risks (including cyber & bio risks),Logging all inputs and outputs from internal use and retaining them for at least 30 days,Analyzing all internal model interactions for abnormal activity, including harmful use or unexpected attempts by AI systems to take real-world actions,Live monitoring and automated editing/resampling of suspicious outputs			

Safety Practices, Frameworks, and Teams (9 Questions)

Question Title	Available options	OpenAI	Z.ai	Anthropic	Google Deepmind	Meta
When you released your latest flagship model, did you release the same model version that the final round of safety (framework) evaluations were conducted on? (Select one)	▪ Yes – we released the same model version. ▪ No – we further modified the model but explicitly mentioned and described all further changes in the model documentation. ▪ No – further modifications are not described explicitly in the model documentation.	No – we further modified the model but explicitly mentioned and described all further changes in the model documentation. There were very minor fine-tuning changes between the model version we tested and the one that was released, and this was not mentioned in the model card summary for 3.1 Pro. We add more detail in our full FSF reports, e.g. the one we released with 3.0 Pro, which satisfies (2) above. We consider 3.0 Pro the last time we released a "new" flagship model for FSF purposes.	Yes – we released the same model version.	Yes – we released the same model version.	No – we further modified the model but explicitly mentioned and described all further changes in the model documentation. There were very minor fine-tuning changes between the model version we tested and the one that was released, and this was not mentioned in the model card summary for 3.1 Pro. We add more detail in our full FSF reports, e.g. the one we released with 3.0 Pro, which satisfies (2) above. We consider 3.0 Pro the last time we released a "new" flagship model for FSF purposes.	Yes – we released the same model version.
If your company has one or more teams focused primarily on technical AI safety research, please provide more information about the team(s) below. *By technical AI safety teams, we are referring to teams researching topics such as scalable oversight, dangerous capability evaluations, mechanistic interpretability, AI control, alignment evaluations, risk-modeling, etc. Please use separate paragraphs for listing multiple teams.*	1) Team name (& website URL if available) 2) Mission and scope – Briefly describe the team's focus. Please distinguish between: ▪ immediate product safety (e.g., RLHF, jailbreak prevention, safety classifiers), and ▪ forward-looking/fundamental research (e.g., model organisms of misalignment, mechanistic interpretability) 3) Technical FTEs – Approximate number of full-time equivalent technical staff (researchers and research engineers). Please count each individual only once, based on their primary team."	We have multiple teams focused primarily on technical AI safety research, led by Johannes Heidecke (Safety Systems) and Mia Glaese (Alignment). Subteams and projects include: ▪ Preparedness ▪ Mechanistic interpretability ▪ CoT interpretability ▪ Automating Alignment ▪ Safety oversight & control ▪ Dangerous capability evaluations ▪ Alignment evaluations ▪ Faithfulness & anti-scheming	▪ **Team name:** Safety Center ▪ **Mission and scope:** mmediate product safety, including jailbreak prevention, safety classifiers, online safety controller etc. ▪ **Technical FTEs:**10+	Aligned with our mission and origin as a safety research lab, we have multiple teams working on AI safety research including alignment science (https://alignment.anthropic.com/), interpretability (https://www.anthropic.com/research#interpretability), frontier red team (http://red.anthropic.com), safeguards (https://alignment.anthropic.com/2025/introducing-safeguards-research-team/ https://www.anthropic.com/news/building-safeguards-for-claude) and more.		MSL Preparedness & Red Teaming & Alignment Team, AI Security Team (https://ai.meta.com/static-resource/muse-spark-safety-and-preparedness-report/).
Does your organization have a formal, written policy that requires notifying external authorities when safety testing determines a model exceeds your organization's "unacceptable-risk" threshold (i.e., a risk-level that bars deployment under your own safety framework), even if the model will not be released? (Select option that best describes your policy)	1) No policy – there is no written requirement to notify any external body. 2) Regulator-only notification – the policy mandates prompt disclosure to a competent regulatory, or supervisory authority. 3) Regulator + public transparency – as in option 2 **and** the policy provides for a public statement or summary once doing so will not exacerbate security risks. Other (please briefly describe):"	1) No policy – there is no written requirement to notify any external body.	Regulator-only notification – the policy mandates prompt disclosure to a competent regulatory, or supervisory authority.		Other (please briefly describe): As outlined in paragraph 5.2 of our Frontier Safety Framework: If we assess that a model has reached a CCL that poses an unmitigated and material risk to overall public safety, we aim to share relevant information with appropriate government authorities where it will facilitate safety of frontier AI. Where appropriate, and subject to adequate confidentiality and security measures and considerations around proprietary and sensitive information."	

Table continues on next page

Question Title	Available options	OpenAI	Z.ai	Anthropic	Google Deepmind	Meta
For companies that signed the ""Frontier AI Safety Commitments"" at the AI Seoul Summit in 2024, and those that strive to implement equivalent safety frameworks: *Which of the levels below best describes the status of your Safety Framework? Please indicate the *highest* option below that accurately describes your current state.*	▪ No official Safety Framework published (yet). ▪ Published & Implementation in progress ▪ Published & substantially implemented – Most discrete policies, processes, or technical safeguards described in the policy are fully implemented and operational. Please briefly assert which elements have not been implemented as described yet and the expected timeline for implementation: ▪ Published & fully implemented – All discrete policies, processes, or technical safeguards described in the policy are fully implemented and operational."	Published & Implementation in progress	No official Safety Framework published (yet).	Published & substantially implemented – Most discrete policies, processes, or technical safeguards described in the policy are fully implemented and operational. Please briefly assert which elements have not been implemented as described yet, the expected timeline for implementation, and any other information related to implementation progress:	Published & fully implemented – All discrete policies, processes, or technical safeguards described in the policy are fully implemented and operational. Please feel free to include any links or additional information related to implementation: We have recently published FSF 3.1 and all discrete policies, processes, and technical safeguards as required by the policy and the risks we have detected are fully implemented and operational. For our framework, see here: https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/strengthening-our-frontier-safety-framework/frontier-safety-framework_3-1.pdf For a description of the tests carried out on our last release of a new model, see here: https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-1-Pro-Model-Card.pdf	Published & substantially implemented – Most discrete policies, processes, or technical safeguards described in the policy are fully implemented and operational. Please briefly assert which elements have not been implemented as described yet, the expected timeline for implementation, and any other information related to implementation progress:

Table continues on next page

Question Title	Available options	OpenAI	Z.ai	Anthropic	Google Deepmind	Meta
Do you have a plan for ensuring that the AGI you're trying to build will remain controllable, safe and beneficial?	▪ No ▪ No, but we're working on it ▪ Yes, internally. (Please briefly explain why you have not published it)	Yes, publicly shared here (please provide URL): Our mission is to ensure that artificial general intelligence benefits all of humanity. For more on our approach to ensuring that AGI remains controllable and safe, see https://openai.com/safety/how-we-think-about-safety-alignment/. As part of our commitment to the broader safety ecosystem, we have accelerated our sharing of safety research work in progress, including via our Alignment Blog (https://alignment.openai.com), launched in December. Work detailed there includes research on interpreting black box reward models (https://alignment.openai.com/argo/), training agents to self-report misbehavior (https://alignment.openai.com/self-incrimination/), and discovering unknown AI misalignments in real-world usage (https://alignment.openai.com/ai-discovered-unknowns/). We are also open sourcing relevant work, including our monitorability evaluations (https://alignment.openai.com/monitorability-evals/). To complement this work, as part of OpenAI's recently concluded recapitalization (https://openai.com/index/built-to-benefit-everyone/), our non-profit, now called the OpenAI Foundation, has made an initial $25 billion commitment to invest in two areas: Health and curing disease, and technical solutions to AI resilience. On March 24, 2026, the Foundation publicly committed (https://openai.com/index/update-on-the-openai-foundation/) to invest at least the first $1 billion of this total within the upcoming 12 months, with AI resilience as a core focus area, led by OpenAI cofounder Wojciech Zaremba, who has moved to the foundation as its Head of AI Resilience. This includes support for independent testing and evaluations, new and stronger industry standards (https://www.commonsensemedia.org/press-releases/common-sense-media-launches-youth-ai-safety-institute), and foundational safety research. "	Yes, internally. (Please briefly explain why you have not published it) This field is in a phase of rapid development, and there are still many uncertainties surrounding the relevant solutions.	Yes, internally. (Please briefly explain why you have not published it) https://trust.anthropic.com/resources?s=gi5v45ke7aezh7b04e82s&name=anthropic-frontier-compliance-framework-[feb-2026-]		Yes, publicly shared here (please provide URL): https://ai.meta.com/static-resource/Meta_Advanced-AI-Scaling-Framework-v2

Table continues on next page

Question Title	Available options	OpenAI	Z.ai	Anthropic	Google Deepmind	Meta
Which of the following elements of an AI emergency response capability has your organization implemented? (Select all that apply)	▪ Maintained and tested technical capability to rapidly roll back a deployed model to a previous version globally (within 12h). Successfully tested rapid full model rollback including internal deployments within the last 12 months. ▪ Maintained and tested technical capability to rapidly tighten model safeguards and restrict specific capabilities (e.g. web-browsing) globally. Successfully tested rapid throttling or capability-restriction including internal deployments within the last 12 months. ▪ Conducted at least one full live emergency response drill/ simulation in the past 12 months. ▪ Created a formal, documented emergency response plan for AI safety incidents with threshold for triggering emergency response, a named incident commander and a 24×7 duty roster. ▪ Established a risk-domain-specific (e.g. bio, cyber) 24-hour communication protocol and points of contact with relevant government agencies. ▪ None of the above ▪ Other: Please use this text-field to share URLs to relevant documentation or to clarify specific responses"	Other: Please use this text-field to share URLs to relevant documentation or to clarify specific responses OpenAI has developed and continues to improve incident response programs across key areas of its operations, including by improving and iterating on our AI safety incident-specific protocols that are tailored to our operations and technology. Our goal is to respond to incidents in a rapid, coordinated way. Our response capabilities include: ▪ Technical Controls for Rapid Mitigation: We maintain the ability to rapidly roll back model deployments globally and to apply restrictions on model functionalities (such as tool use or capability throttling) in response to emergent risks. The roll back mechanism was successfully utilized last year in response to our finding that a GPT-4o model update was overly flattering or agreeable (see Sycophancy in GPT-4o: what happened and what we're doing about it, https://openai.com/index/sycophancy-in-gpt-4o/). ▪ Incident Response Planning and Structure: OpenAI has formal incident response plans for key areas of operations, including AI safety incident-specific protocols. Our response activities include escalation thresholds and mechanisms as well as incident response functions, such as response leads and as on-call rotations across functions to support implementation of response activity. We maintain close coordination across research, engineering, safety, legal, communications and policy teams, and have integrated lessons learned into our formal plans. ▪ As part of our commitment to continuous improvement, we continue to refine our incident response capabilities, including robust playbooks for rapid-response. These efforts are integral to our broader model governance and safety assurance frameworks.	Maintained and tested technical capability to rapidly roll back a deployed model to a previous version globally (within 12h). Successfully tested rapid full model rollback including internal deployments within the last 12 months.,Maintained and tested technical capability to rapidly tighten model safeguards and restrict specific capabilities (e.g. web-browsing) globally. Successfully tested rapid throttling or capability-restriction including internal deployments within the last 12 months.,Conducted at least one full live emergency response drill/ simulation in the past 12 months.,Created a formal, documented emergency response plan for AI safety incidents with threshold for triggering emergency response, a named incident commander and a 24 × 7 duty roster.,Established a risk-domain-specific (e.g. bio, cyber) 24-hour communication protocol and points of contact with relevant government agencies.	Other: Please use this text-field to share URLs to relevant documentation or to clarify specific responses Please see our RSP, Frontier Compliance Framework and transparency hub for more. https://cdn.sanity.io/files/4zrzovbb/website/28c6241900d90410628a8a2003a5572faae4365a.pdf https://trust.anthropic.com/resources?s=gi5v45ke7aezh7b04e82s&name=anthropic-frontier-compliance-framework-[feb-2026-] https://www.anthropic.com/transparency		

Table continues on next page

Question Title	Available options	OpenAI	Z.ai	Anthropic	Google Deepmind	Meta
Does your company agree with the following principles for promoting legible and faithful reasoning in advanced AI systems to ensure AI remains safe and controllable? (Select all statements you support) *Leading AI companies should:*	▪ **Ensure Human-Legible Reasoning** - AI models should reason in ways that are accessible and understandable to humans. Developers should avoid opaque reasoning methods. [No.] ▪ **Avoid Optimization That Encourages Obfuscation** - Developers should exercise caution when applying optimization pressures to model reasoning, especially when removing 'undesired reasoning', to prevent fostering deceptive behavior. ▪ **Disclose Optimization Pressures on Reasoning** - Companies should transparently report the optimization pressures and training methods applied to model reasoning, particularly when removing 'undesired reasoning.' ▪ None of the above	*Avoid Optimization That Encourages Obfuscation* - Developers should exercise caution when applying optimization pressures to model reasoning, especially when removing 'undesired reasoning', to prevent fostering deceptive behavior. We've published research (https://arxiv.org/abs/2503.11926) and joined a broader working paper (https://arxiv.org/abs/2507.11473) urging against optimizing on chains of thought: As we noted in the GPT-5 system card (https://cdn.openai.com/gpt-5-system-card.pdf), "our commitment to keep our reasoning models' CoTs as monitorable as possible (i.e., as faithful and legible as possible) allows us to conduct studies into our reasoning models' behavior by monitoring their CoTs." We've introduced new research and evaluations focused on CoT Monitorability (https://openai.com/index/reasoning-models-chain-of-thought-controllability/)that we now include in our system cards. In early May, we published an update (https://alignment.openai.com/accidental-cot-grading/) describing a case in which we found limited accidental CoT grading in some released models, fixed the affected reward pathways, and found no clear evidence that monitorability degraded.	***Ensure Human-Legible Reasoning*** - AI models should reason in ways that are accessible and understandable to humans. Developers should avoid opaque reasoning methods. *Avoid Optimization That Encourages Obfuscation* - Developers should exercise caution when applying optimization pressures to model reasoning, especially when removing 'undesired reasoning', to prevent fostering deceptive behavior.			
Task-Specific Fine-Tuning (TSFT) involves training a model to excel at potentially dangerous tasks (e.g., designing biological agents, cyber attacks). *Before releasing your current frontier model, which statement best describes your TSFT safety testing? (Select one)*	▪ None – no TSFT safety testing performed (skips follow-up). ▪ Partial – TSFT performed on ≤ 2 high-risk domains (choose below). ▪ Comprehensive – TSFT performed on ≥ 3 high-risk domains (choose below)."	None – no TSFT safety testing performed (skips follow-up). None for gpt-5.5. We evaluated helpful-only models, which we believe is appropriate for the threat model of misuse for models made available via our platform and whose weights we do not release, as is codified in our Preparedness Framework. Note that we did task-specific fine tuning on biological and cyber capabilities for gpt-oss and published a paper with our findings, Estimating worst case frontier risks of open weight LLMs (https://openai.com/index/estimating-worst-case-frontier-risks-of-open-weight-llms/)	Partial – TSFT performed on ≤ 2 high-risk domains (choose below).	None – no TSFT safety testing performed (skips follow-up).		
If you selected 'Partial' or 'Comprehensive' on the previous question, Please tick the risk-domains tested with TSFT.	▪ Biological ▪ Persuasion ▪ Chemical ▪ Deceptive alignment / Autonomy ▪ Cyber-offense ▪ Other (please specify):		Deceptive alignment / Autonomy			

Table continues on next page

Question Title	Available options	OpenAI	Z.ai	Anthropic	Google Deepmind	Meta
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. You may also share additional information about your company's policies.		**Additional information:** **Mental health:** ▪ We publicly introduced new "dynamic" predeployment evaluations for mental health, emotional reliance and self-harm in the GPT 5.3-Instant system card (https://deploymentsafety.openai.com/gpt-5-3-instant/disallowed-content/tbl-2), and these are now part of our standard suite of launch evaluations. Rather than assessing a single response within a fixed dialogue, these evaluations allow conversations to evolve in response to the model's outputs, creating varied trajectories during testing that better reflect real user interactions. This approach helps identify potential issues that may only emerge over the course of long exchanges and provides an even more rigorous test than prior static multi-turn methods. By utilizing realistic, yet adversarial user simulations, these evaluations have enabled continued improvements in safety performance, particularly in areas where earlier evaluation frameworks had reached saturation. We continue to report these metrics in our system cards for frontier model deployments. ▪ We consult our Expert Council on Well-Being and AI (https://openai.com/index/expert-council-on-well-being-and-ai/) on mental health, parental controls and other topics related to well-being. ▪ Our program of externally funding independent research on AI and mental health (https://openai.com/index/ai-mental-health-research-grants/), with a grant budget of $2 million received more than 1,000 applications. Funded projects are currently underway. **Public Policy:** We have published policy blueprints describing OpenAI's views and policy proposals on (1) how the US can maximize AI's benefits, bolster national security, and drive economic growth (https://openai.com/global-affairs/openais-economic-blueprint/) (2) early policy proposals for superintelligence, designed to expand opportunity, share prosperity, and build resilient institutions (https://openai.com/index/industrial-policy-for-the-intelligence-age/), (3) how to combat and prevent AI-enabled sexual exploitation of children (https://openai.com/index/introducing-child-safety-blueprint/) and (4) how AI should work to protect teens (https://openai.com/index/introducing-the-teen-safety-blueprint/). **Security:** Security: Below, we include some additional information about our security work that we believe may be useful context for evaluators considering our overall posture and approach. For additional technical detail on our security measures for AI see: Security on the path to AGI **Third party collaboration on security:** OpenAI maintains a bug bounty program through BugCrowd (https://bugcrowd.com/openai), and welcomes responsible disclosures from third parties via our coordinated vulnerability disclosure policy (https://openai.com/policies/coordinated-vulnerability-disclosure-policy/). In addition, OpenAI runs a Cybersecurity Grant Program to support research and development focused on protecting AI systems and infrastructure. This program encourages and funds initiatives that help identify and address vulnerabilities, ensuring the safe deployment of AI technologies.				

FLI AI Safety Index

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

July 2026