. **Introduction:**
Thank you for participating in the **FLI AI Safety Index 2025 Survey**. This survey is designed to allow your company to showcase 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.

Methodological Continuity:
We intentionally retain the same question set from the last edition of the Index to facilitate comparisons over time. Please respond based on your company’s current status at the time of submission.

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

Contact:
If you have questions about a survey item or need clarification on what is being asked, you may contact sabina@futureoflife.org.

We appreciate your time and effort in providing thorough answers.

. **Whistleblowing policies page 1/2** (1 Question)

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.

Q1. Does your company have a WB policy & function covering frontier AI-focused staff?
Is this policy publicly accessible without login credentials?

- ○ 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:

  Please see “post deployment monitoring” in our transparency hub. We expect to share more publicly in the near future. https://www.anthropic.com/transparency/voluntary-commitments

- ○ Public WB policy - Please provide URL here:

. **Whistleblowing policies page 2/2** (15 Questions)

Q2. Who is formally designated with primary responsibility for overseeing the whistleblowing function and ensuring reports are properly addressed?

- ○ Board/Audit Committee
- ○ Executive management
- ○ Compliance/Legal department
- ○ HR department

- ○ Other (Please also specify whom this role reports to):

Q3. 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.

Q4. 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

Q5. 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

Q6. 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

- [x] Suppliers and vendors with access to company systems

Q7. To which of the following individuals or entities can whistleblowers submit reports according to your policy? (Select all that apply)

- [ ] 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): _____

Q8. For former employees and contractors, indicate any policy limitations compared with current employees. (Select all limitations that apply)

	A	B
	Former employees	Contractors
Limited Reporting Channels	○	○
Limited Reportable Issues	○	○
Limited Retaliation Protection	○	○
No Limitations	◉	◉

Q9. 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

Q10. 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)

Q11. If “Limited”, under which circumstances is external disclosure protected?

This question was not displayed to the respondent.

Q12. Which mechanisms ensure that your whistleblowing function has access to adequate (technical) expertise to investigate reports? (Select all that apply)

- ☐ 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):

Q13. 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

Q14. 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

Q15. 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

Q16. Which anti-retaliation provisions are explicitly detailed in your whistleblowing policy? (Select all that apply)

- [x] Defined disciplinary consequences for individuals who retaliate against whistleblowers (e.g., termination, demotion, or other concrete penalties - not just general statements prohibiting retaliation)
- [x] 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

. 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.

. **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 4 Opus
DeepSeek - R1
Google Deepmind - Gemini 2.5 Pro
Meta - Llama 4 Maverick
OpenAI - o3
xAI - Grok3
Zhipu AI - GLM4 Plus

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

Q17. 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?

- ○ 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):

```
Please see our system cards (library,
Claude Opus 4) and transparency hub
for information on our external
testing

https://docs.claude.com/en/resources/o
verview

https://www-
cdn.anthropic.com/6d8a8055020700718b0c
49369f60816ba2a7c285.pdf

https://www.anthropic.com/transparency
/voluntary-commitments
```

Q18. 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

Q19. 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

Q20. 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)

- ☐ 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:

Q21. 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.

Q22. Does your organization log and retain the model interactions of external evaluators during pre-deployment testing?

- ◯ Yes - Inputs and outputs are logged and retained.
- ◯ No - Inputs and outputs are neither logged nor retained, protecting evaluator IP.
- ◯ Other (please describe):

. 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.

. **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.

Q23. 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):

Q24. 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):

Q25. 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):

. 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.

Re: 25 we have nuanced rigorous approach to safeguards- each of these depends on product surface, classifier and harm type, and use case.

. **Safety practices, frameworks, and teams** (9 Questions)

Q26. 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.

Q27. 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.

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.

Q28. 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):

U.S. Government notice when model requires ASL-3+ safeguards; see our RSP for more

Q29. 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.

Q30. 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): https://www-cdn.anthropic.com/872c653b2d0501d6ab44cf87f43e1dc4853e4d37.pdf

Q31. 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

- [x] Other: Please use this text-field to share URLs to relevant documentation or to clarify specific responses

Please see our RSP and transparency hub for more

Q32. 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.
- [ ] **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

Q33. **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).

Q34. If you selected 'Partial' or 'Comprehensive' on the previous question, Please tick the risk-domains tested with TSFT.

This question was not displayed to the respondent.

Q44. 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.