Source: https://transparency.oecd.ai/reports/bed824e5-b9af-44ba-9bbf-630cdfa9029b

Report

Organization: **Anthropic**

anthropic.com

Publication date: Apr 22, 2025, 09:01 AM PDT

Reporting period: 2025

Section 1 - Risk identification and evaluation

a. How does your organization define and/or classify different types of risks related to AI, such as unreasonable risks?

Please see the “Risk Identification” section of our Transparency Hub under Voluntary Commitments > Risk Assessment and Mitigation.

b. What practices does your organization use to identify and evaluate risks such as vulnerabilities, incidents, emerging risks and misuse, throughout the AI lifecycle?

Please see the “Internal and External Risk Assessments”, “Post-Deployment Monitoring”, and “Information Sharing on Risks and Threats” sections of our Transparency Hub under Voluntary Commitments > Risk Assessment and Mitigation.

c. Describe how your organization conducts testing (e.g., red-teaming) to evaluate the model’s/system’s fitness for moving beyond the development stage?

Please see the “Internal and External Risk Assessments” section of our Transparency Hub under Voluntary Commitments > Risk Assessment and Mitigation.

d. Does your organization use incident reports, including reports shared by other organizations, to help identify risks?

Yes

e. Are quantitative and/or qualitative risk evaluation metrics used and if yes, with what caveats? Does your organization make vulnerability and incident reporting mechanisms accessible to a diverse set of stakeholders? Does your organization have incentive programs for the responsible disclosure of risks, incidents and vulnerabilities?

For information about our risk evaluation methods and metrics, please see the “Risk Identification” and “Internal and External Risk Assessments” sections of our Transparency Hub under Voluntary Commitments > Risk Assessment and Mitigation.

For information about our vulnerability and incident reporting as well as incentive programs such as bug bounties, please see the “Post-Deployment Monitoring” section of our Transparency Hub under Voluntary Commitments > Risk Assessment and Mitigation.

f. Is external independent expertise leveraged for the identification, assessment, and evaluation of risks and if yes, how? Does your organization have mechanisms to receive reports of risks, incidents or vulnerabilities by third parties?

For information about external evaluations, please see the “Internal and External Risk Assessments” section of our Transparency Hub under Voluntary Commitments > Risk Assessment and Mitigation.

For information about our vulnerability and incident reporting, please see the “Post-Deployment Monitoring” section of our Transparency Hub under Voluntary Commitments > Risk Assessment and Mitigation.

g. Does your organization contribute to the development of and/or use international technical standards or best practices for the identification, assessment, and evaluation of risks?

Please see the “Advancements of Global Technical Standards” section of our Transparency Hub under Voluntary Commitments > Public Awareness.

h. How does your organization collaborate with relevant stakeholders across sectors to assess and adopt risk mitigation measures to address risks, in particular systemic risks?

Please see the “Information Sharing on Risks and Threats” section of our Transparency Hub under Voluntary Commitments > Risk Assessment and Mitigation.

Section 2 - Risk management and information security

a. What steps does your organization take to address risks and vulnerabilities across the AI lifecycle?

Please see the “Risk Assessment and Mitigation” section of our Transparency Hub under Voluntary Commitments.

b. How do testing measures inform actions to address identified risks?

Please see the “Internal and External Risk Assessments” section of our Transparency Hub under Voluntary Commitments > Risk Assessment and Mitigation.

c. When does testing take place in secure environments, if at all, and if it does, how?

Please see the “Security During External Testing” section of our Transparency Hub under Voluntary Commitments > Security and Privacy.

d. How does your organization promote data quality and mitigate risks of harmful bias, including in training and data collection processes?

Please see the “Model Report” section of our Transparency Hub for the latest information about training data and bias evaluation results.

e. How does your organization protect intellectual property, including copyright-protected content?

Please see the “Model Report” section of our Transparency Hub and the “Training Data & Process” section of our linked System Card for Claude 3.7 Sonnet.

f. How does your organization protect privacy? How does your organization guard against systems divulging confidential or sensitive data?

Please see the “Protections for Personal Data” section of our Transparency Hub under Voluntary Commitments > Security and Privacy and visit our Privacy Center for more information.

g. How does your organization implement AI-specific information security practices pertaining to operational and cyber/physical security?

- i. How does your organization assess cybersecurity risks and implement policies to enhance the cybersecurity of advanced AI systems?
- ii. How does your organization protect against security risks the most valuable IP and trade secrets, for example by limiting access to proprietary and unreleased model weights? What measures are in place to ensure the storage of and work with model weights, algorithms, servers, datasets, or other relevant elements are managed in an appropriately secure environment, with limited access controls in place?
- iii. What is your organization's vulnerability management process? Does your organization take actions to address identified risks and vulnerabilities, including in collaboration with other stakeholders?
- iv. How often are security measures reviewed?
- v. Does your organization have an insider threat detection program?

Please see the "Cybersecurity and Insider Threat Safeguards" section of our Transparency Hub under Voluntary Commitments > Security and Privacy and the "Risk Assessment and Mitigation" section under Voluntary Commitments

h. How does your organization address vulnerabilities, incidents, emerging risks?

Please see the "Post-Deployment Monitoring" section of our Transparency Hub under Voluntary Commitments > Risk Assessment and Mitigation.

Section 3 - Transparency reporting on advanced AI systems

a. Does your organization publish clear and understandable reports and/or technical documentation related to the capabilities, limitations, and domains of appropriate and inappropriate use of advanced AI systems?

- i. How often are such reports usually updated?
- ii. How are new significant releases reflected in such reports?
- iii. Which of the following information is included in your organization's publicly available documentation: details and results of the evaluations conducted for potential safety, security, and societal risks including risks to the enjoyment of human rights; assessments of the model's or system's effects and risks to safety and society (such as those related to harmful bias, discrimination, threats to protection of privacy or personal data, fairness); results of red-teaming or other testing conducted to evaluate the model's/system's fitness for moving beyond the development stage; capacities of a model/system and significant limitations in performance with implications for appropriate use domains; other technical documentation and instructions for use if relevant.

i. Please see our Transparency Hub which includes a link to our System Card for Claude 3.7 Sonnet.

ii. We publish model/system cards or addendums with each model release and update our Transparency Hub accordingly.

iii. All of the information listed is included in publicly available documentation.

b. How does your organization share information with a diverse set of stakeholders (other organizations, governments, civil society and academia, etc.) regarding the outcome of evaluations of risks and impacts related to an advanced AI system?

Please see the "Risk Assessment and Mitigation" and "Public Awareness" sections of our Transparency Hub under Voluntary Commitments.

c. Does your organization disclose privacy policies addressing the use of personal data, user prompts, and/or the outputs of advanced AI systems?

Please see the "Protections for Personal Data" section of our Transparency Hub under Voluntary Commitments > Security and Privacy. Further, our Privacy Center discloses more details regarding our privacy policy.

d. Does your organization provide information about the sources of data used for the training of advanced AI systems, as appropriate, including information related to the sourcing of data annotation and enrichment?

Please see the “Model Report” section of our Transparency Hub for information about Training Data.

e. Does your organization demonstrate transparency related to advanced AI systems through any other methods?

Please see our Transparency Hub, including in particular the “Public Report on AI Systems” section under Voluntary Commitments > Public Awareness.

Section 4 - Organizational governance, incident management and transparency

a. How has AI risk management been embedded in your organization governance framework? When and under what circumstances are policies updated?

Anthropic’s legal status as a public benefit corporation aligns our corporate governance with our mission of developing and maintaining advanced AI for the long-term benefit of humanity. As a part of our mission, we build frontier LLMs in order to conduct empirical safety research and to deploy commercial models that are beneficial and useful to society.

Please also see the “Responsible Scaling Policy” section of our Transparency Hub under Voluntary Commitments > Risk Assessment and Mitigation.

b. Are relevant staff trained on your organization’s governance policies and risk management practices? If so, how?

Yes, all staff receive security and privacy training including reading and acknowledging security and privacy policies at hire and on an annual basis thereafter. Additionally, all staff are educated on the Responsible Scaling Policy during onboarding to the company, with regular updates during our company-wide internal meetings.

c. Does your organization communicate its risk management policies and practices with users and/or the public? If so, how?

Please see the “Responsible Scaling Policy” section of our Transparency Hub under Voluntary Commitments > Risk Assessment and Mitigation, including the link to the Responsible Scaling Policy itself.

d. Are steps taken to address reported incidents documented and maintained internally? If so, how?

Please see the “Post-Deployment Monitoring” section of our Transparency Hub under Voluntary Commitments > Risk Assessment and Mitigation.

e. How does your organization share relevant information about vulnerabilities, incidents, emerging risks, and misuse with others?

Please see the “Information Sharing on Risks and Threats” section of our Transparency Hub under Voluntary Commitments > Risk Assessment and Mitigation.

f. Does your organization share information, as appropriate, with relevant other stakeholders regarding advanced AI system incidents? If so, how? Does your organization share and report incident-related information publicly?

Please see the “Information Sharing on Risks and Threats” section of our Transparency Hub under Voluntary Commitments > Risk Assessment and Mitigation.

Specifically, as a founding member of the Frontier Model Forum (FMF), we have signed an agreement designed to facilitate information-sharing about threats, vulnerabilities, and capability advances unique to frontier AI.

g. How does your organization share research and best practices on addressing or managing risk?

Please see the “Information Sharing on Risks and Threats” section of our Transparency Hub under Voluntary Commitments > Risk Assessment and Mitigation and the “Public Awareness” section under Voluntary Commitments.

h. Does your organization use international technical standards or best practices for AI risk management and governance policies?

Please see the "Advancements of Global Technical Standards" section of our Transparency Hub under Voluntary Commitments > Public Awareness, the "Responsible Scaling Policy" section under under Voluntary Commitments > Risk Assessment and Mitigation, and visit our Trust Center to learn more about our compliance and security certifications.

Section 5 - Content authentication & provenance mechanisms

a. What mechanisms, if any, does your organization put in place to allow users, where possible and appropriate, to know when they are interacting with an advanced AI system developed by your organization?

Anthropic employs multiple strategies, including but not limited to disclaimers, to ensure users are aware that they are interacting with an advanced AI system. We want our users to understand that Claude is a generative AI assistant, and this is highlighted throughout our marketing and onboarding. Examples of this transparency include reminders to users who are signing up for Claude.ai accounts that they will be interacting with an AI assistant, disclaimers in our Consumer Terms of Service, in-conversation reminders that Claude can make mistakes, and documentation for developers about the nature of the AI and best practices for implementation.

b. Does your organization use content provenance detection, labeling or watermarking mechanisms that enable users to identify content generated by advanced AI systems? If yes, how? Does your organization use international technical standards or best practices when developing or implementing content provenance?

Please see the "Transparency of AI Generation" section of our Transparency Hub.

Section 6 - Research & investment to advance AI safety & mitigate societal risks

a. How does your organization advance research and investment related to the following: security, safety, bias and disinformation, fairness, explainability and interpretability, transparency, robustness, and/or trustworthiness of advanced AI systems?

Please see the “Risk Identification", “Internal and External Risk Assessments”, “Advancements of Global Technical Standards,” “Public Report on AI Systems”, and “Public Benefit Research and Support” sections of our Transparency Hub under Voluntary Commitments.

b. How does your organization collaborate on and invest in research to advance the state of content authentication and provenance?

Please see the “Transparency of AI Generation” section of our Transparency Hub under Voluntary Commitments > Public Awareness.

c. Does your organization participate in projects, collaborations, and investments in research that support the advancement of AI safety, security, and trustworthiness, as well as risk evaluation and mitigation tools?

Please see the “Risk Assessment and Mitigation” section of our Transparency Hub under Voluntary Commitments.

d. What research or investment is your organization pursuing to minimize socio-economic and/or environmental risks from AI?

Please see the “Public Benefit Research and Support” and the “Economic Impact Research” sections of our Transparency Hub under Voluntary Commitments > Societal Impact.

Section 7 - Advancing human and global interests

a. What research or investment is your organization pursuing to maximize socio-economic and environmental benefits from AI? Please provide examples.

Please see the “Public Benefit Research and Support” and the “Economic Impact Research” sections of our Transparency Hub under Voluntary Commitments > Societal Impact.

b. Does your organization support any digital literacy, education or training initiatives to improve user awareness and/or help people understand the nature, capabilities, limitations and impacts of advanced AI systems? Please provide examples.

Please see the “AI Education and Professional Development” section of our Transparency Hub under Voluntary Commitments > Societal Impact.

c. Does your organization prioritize AI projects for responsible stewardship of trustworthy and human-centric AI in support of the UN Sustainable Development Goals? Please provide examples.

Please see the “Public Benefit Research and Support” section of our Transparency Hub under Voluntary Commitments > Societal Impact., which highlights how we use Constitutional AI in an effort to better align our models with human values.

d. Does your organization collaborate with civil society and community groups to identify and develop AI solutions in support of the UN Sustainable Development Goals and to address the world’s greatest challenges? Please provide examples.

Please see the “Public Benefit Research and Support” section of our Transparency Hub under Voluntary Commitments > Societal Impact.