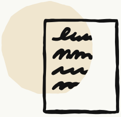

Anthropic’s Transparency Hub: Voluntary Commitments

A look at Anthropic's key processes, programs,
and practices for responsible AI development.

Executive Summary

Last updated Nov 19, 2024

Below is information about how we are meeting and working towards our voluntary commitments. Our experience with multiple voluntary frameworks has revealed consistent themes, as well as considerable overlap in their core requirements around safety, security, and responsible development. We are providing an overview organized by key areas of focus. We welcome feedback from the AI community and policymakers to inform our future work.

Risk Assessment and Mitigation

Responsible Scaling Policy

In September 2023, we published the first version of our Responsible Scaling Policy (RSP), our framework for managing potential catastrophic risks from models.

The policy is centered around implementing safeguards which are proportional to the identified risks. As AI models become more powerful, they require stronger protections. When models reach certain capability thresholds, we will implement additional safeguards around security and deployment.

The RSP is designed to evolve as our understanding of AI risks improves, while maintaining this fundamental commitment to safety. It serves both as our internal guidebook and as a model for industry-wide safety standards.

Related Commitments: G7 Hiroshima Process International Code of Conduct; AI Seoul Summit's Frontier AI Safety Commitments; Seoul AI Business Pledge

Risk Identification

Anthropic works to identify a wide spectrum of potential risks from AI systems:

- For risks addressed in our Responsible Scaling Policy, we have identified Capability Thresholds, which we think would require stronger safeguards than our current baseline measures provide. (In other words, we think that models with such capabilities, if stored under our current safeguards, would present intolerable levels of risk.) We have adopted capability thresholds for CBRN

weapons and autonomous AI research and development.
- We also study and assess risks in other domains, including cybersecurity; autonomous capabilities; societal impacts like representation and discrimination; and child safety and election integrity.

This system is dynamic and evolving. We also regularly update our Usage Policy to reflect new insights into how our models are being used and adjust our risk identification and assessment strategies accordingly.

Related Commitments: AI Seoul Summit's Frontier AI Safety Commitments

Internal and External Risk Assessments

Anthropic employs a multi-faceted approach to assessing and mitigating catastrophic and non-catastrophic risks across the AI lifecycle. For example, we may employ the following techniques:

1. **Regular evaluations:** We conduct systematic evaluations at defined intervals to detect warning signs of increased catastrophic risks.
2. **Threat modeling:** We collaborate with external experts to develop detailed threat models, particularly in high-risk areas as outlined in our RSP.
3. **Red team testing:** We employ both internal and external red teaming to proactively identify vulnerabilities and potential misuse scenarios. This includes testing for issues like deception, jailbreaking, emergent capabilities, as well as potential misuse scenarios for risks covered under our Usage Policy, such as engaging in fraud or inciting violence.
4. **Expert consultations:** We integrate feedback from external subject matter experts to ensure our risk identification processes are robust.
5. **External evaluations:** We work with independent organizations like the UK AI Safety Institute (UK AISI), the US AI Safety Institute (US AISI), and Model Evaluation and Threat Research (METR) to conduct additional testing and evaluation of our models.
6. **Research on emergent risks:** Our research teams actively investigate potential future risks, such as autonomous AI R&D.
7. **Policy vulnerability testing (PVT):** Our Safeguards team conducts in-depth, qualitative testing on a variety of policy topics covered under our Usage Policy.
8. **Pre-deployment testing:** Before releasing new models, we conduct thorough testing to identify potential risks.
9. **External certifications:** Our compliance teams work to achieve accredited certifications that provide independent validation of our risk frameworks. We achieved the ISO/IEC 42001:2023 standard for our AI management system, which is the first international standard outlining requirements for AI governance and helps ensure AI systems are developed and used responsibly.

Related Commitments: G7 Hiroshima Process International Code of Conduct; AI Seoul Summit's Frontier AI Safety Commitments; Seoul AI Business Pledge

Post-Deployment Monitoring

We regularly update our Usage Policy and our classifiers based on how our models are being used in practice. Additionally, Anthropic has also established multiple mechanisms for receiving reports of incidents and vulnerabilities from third parties:

1. **Responsible Disclosure Policy:** We have a publicly accessible Responsible Disclosure Policy on our website with a reporting form for security-related vulnerabilities.
2. **Bug Bounty Program:** We operate two private bug bounty programs through HackerOne: one for identifying model safety issues and one for security vulnerabilities.
3. **Safety Issue Reporting:** Claude.ai and Claude API users can report safety issues, “jailbreaks”, and similar concerns at

usersafety@anthropic.com.

4. **Engagement with Research Community:** We maintain open channels of communication with the broader AI research community, allowing for informal reporting of potential issues or concerns.

Related Commitments: G7 Hiroshima Process International Code of Conduct

Information Sharing on Risks and Threats

1. We are a founding member of the Frontier Model Forum (FMF), an industry organization developing safety research, standards, and evaluations for AI safety and responsibility.
2. We collaborate with government organizations like the UK and US AI Safety Institutes for independent testing.
3. We have launched an initiative to fund third-party evaluations of advanced AI capabilities.
4. We partner with academic researchers to advance the science of AI safety and evaluation.
5. We work with domain experts to improve our risk assessments in specific areas.
6. We engage globally on topics such as child safety and election integrity, collaborating with civil society, industry, and governments to share research and gain insights.

Related Commitments: G7 Hiroshima Process International Code of Conduct; AI Seoul Summit's Frontier AI Safety Commitments; Seoul AI Business Pledge

Security & Privacy

Cybersecurity and Insider Threat Safeguards

Anthropic implements a number of operational and cybersecurity best practices:

1. **Security Controls:** We implement a number of cybersecurity measures and security safeguards, specifically tailored for AI model development.
2. **Third-Party Evaluations:** We engage with independent assessors to evaluate the effectiveness of our security controls and participate in third-party test and evaluation schemes periodically. We make these attestation and compliance artifacts available on our Trust Center.
3. **Regular Threat Modeling:** We perform regular reviews and updates to our threat model considering tactics, techniques, and procedures used by significant threat actors, including nation-states.
4. **AI-Specific Risk Mitigation:** We conduct research on making models more resistant to prompt injection and other adversarial "jailbreaking" techniques. We also use red teaming to evaluate model vulnerabilities and implement mitigations.
5. **Supply Chain Security:** We implement inspection and control measures over our third-party supply chain to mitigate potential risks.
6. **Enterprise Customers:** We offer enterprise-grade security features to ensure customer data is handled safely and securely. These features include SSO, SCIM, audit logs, and role-based permissions. See more in our Help Center.

Related Commitments: G7 Hiroshima Process International Code of Conduct

Security During External Testing

We protect the security of the environment of our models, including during evaluations:

- Models are protected by two-party controls, with explicit per-user access validation and multifactor authentication.
- Internal model evaluations are performed within our own infrastructure, while external evaluations use API access with ‘zero data retention’ settings to prevent content storage.

Related Commitments: G7 Hiroshima Process International Code of Conduct

Protections for Personal Data

We respect privacy rights and comply with relevant data protection laws, including through detailed disclosures about personal data use and processing. By default Anthropic generative models are not trained on any user prompt or output data submitted to us by users or customers, including free users, Claude Pro users, and API customers. We implement technical measures to respect intellectual property rights, including respecting robots.txt.

More details can be found in our Privacy Center, Help Center, and our Privacy Policy.

Related Commitments: G7 Hiroshima Process International Code of Conduct

Public Awareness

Advancements of Global Technical Standards

Anthropic contributes to the development of international technical standards and best practices:

- We collaborated with NIST to support development of their AI Risk Management Framework by sharing insights from our technical safety research for incorporation into the companion playbook.
- We co-founded and are an active member in the Frontier Model Forum (FMF), which, among other aims, seeks to advance AI safety research, standards and evaluations.
- We collaborate with the Cloud Security Alliance (CSA) on the development of controls applicable to the AI industry and assist in the development of diligence efforts that could take place based on those controls.

We are actively contributing to the development of standards for evaluating models and third-party testing, by launching an initiative to fund evaluations developed by third-party organizations that can effectively measure advanced capabilities in AI models and by proposing a third-party testing regime.

Related Commitments: G7 Hiroshima Process International Code of Conduct; Seoul AI Business Pledge

Public Report on AI Systems

Anthropic publishes and maintains detailed information on our models and practices:

1. **Model Cards:** With each new model release, we publish a detailed model card or addendum. These cards provide information about model capabilities and performance across various benchmarks; known limitations and potential risks; results of safety evaluations and red teaming; information on model training; and more.
2. **Responsible Scaling Policy (RSP):** We make public our RSP that outlines our framework for evaluating and mitigating potential catastrophic risks posed by AI systems.
3. **Research:** We regularly publish research on cutting edge safety, interpretability and social impacts.
4. **Usage Policy:** Our Usage Policy is intended to help our users stay safe and help ensure our products and services are being used responsibly.
5. **User Guides:** We publish a suite of reference documents for users to learn more about Claude's capabilities and appropriate uses including a Prompt Library, Prompt Engineering Guidelines, Release Notes, and System Prompt updates.

Related Commitments: G7 Hiroshima Process International Code of Conduct; AI Seoul Summit's Frontier AI Safety Commitments; Seoul AI Business Pledge

Transparency of AI Generation

Claude currently has multimodal input capabilities and text-only output. While watermarking is most commonly applied to image outputs, we continue to work across industry and academia to explore and stay abreast of technological developments in this area.

Related Commitments: G7 Hiroshima Process International Code of Conduct; Seoul AI Business Pledge

Societal Impact

Public Benefit Research and Support

Many of our enterprise customers leverage Claude to increase public health, environmental, and social benefits:

- Educational startups like Juni Learning, which has integrated Claude to help its students achieve academic success, delivering conversational assistance at the level of a true tutor, across a range of subjects like math and critical reading;
- Climate companies like BrainBox AI, whose AI technology partners enabled building operators to reduce energy costs by up to 25% and greenhouse gas emissions by up to 40%; and
- Pfizer, one of the world's premier biopharmaceutical companies, which is using Claude in the discovery of potential treatments for cancer to get breakthroughs to patients faster. With Claude, Pfizer can gather relevant data and scientific content in a fraction of the time, and then use it to assess trends and generate and validate oncology targets, improving the probability of success.

We also have researcher access programs to provide free Claude credits for researchers advancing AI research.

Anthropic also conducts research on election integrity, discriminatory model outputs, socioeconomic impacts of AI, and more. We have created Constitutional AI in an effort to better align our models with human values. Within this spirit, we also ran an experiment and published our findings on “Collective Constitutional AI,” an effort to collect and incorporate a diverse range of human perspectives and ethical stances into a sample model's training, aiming to create a more globally representative and culturally sensitive AI system.

Related Commitments: G7 Hiroshima Process International Code of Conduct; Seoul AI Business Pledge

Economic Impact Research

We launched the Anthropic Economic Index, an initiative aimed at understanding AI's effects on labor markets and the economy over time. The index provides analysis based on millions of anonymized Claude conversations, offering insights into how AI is being incorporated into real-world tasks across the modern economy. As part of this effort, we are open sourcing the dataset underpinning our analysis and inviting economists, policy experts, and researchers to provide input on the Index. We plan to release regular updates to support longitudinal analysis of AI use patterns over time.

Related Commitments: G7 Hiroshima Process International Code of Conduct; Seoul AI Business Pledge

AI Education and Professional Development

Anthropic enables organizations and professionals to learn and work with AI tools:

1. We launched Claude for Enterprise, which helps organizations securely collaborate using internal knowledge in our AI chatbot.
2. Our Prompt Library provides a library of optimized prompts for business and personal tasks.
3. We maintain academic partnerships and created an External Researcher Access program to foster collaboration between industry and academia.

4. Many of our customers are startups and SMEs that use the Claude API to improve their offerings via AI technology. By serving a wide range of industries through our API, we’re facilitating the integration of AI into multiple professions ranging from architecture to administrative support, thereby driving productivity improvements.

Related Commitments: Seoul AI Business Pledge

Democratizing Model Access

Claude is available in over 160 countries. We will continue to invest and grow our efforts to internationalize our products in an inclusive and localized way.

We support the National AI Research Resource (NAIRR), which is an initiative for national infrastructure that connects U.S. researchers to computational, data, software, model and training resources they need to participate in AI research.

We endorsed the CREATE AI Act to authorize the NAIRR and are participating in the NAIRR pilot at the National Science Foundation.

Related Commitments: Seoul AI Business Pledge

Platform Safeguard Commitments

In the following sections, we'll focus specifically on how we address three critical areas that have their own dedicated sets of commitments: Image-Based Sexual Abuse, Election Integrity, and Terrorist and Extremist Content.

Policy Prohibitions

At the foundation of our Safeguards work is our Usage Policy, which strictly prohibits using our products and services for child exploitation, sexually explicit content, activities related to political campaigns or elections, activities associated with terrorist and violent extremist content, and more. We have a suite of tools to enforce our policies and minimize potential harm. These include:

- Advanced classifiers, which are AI-powered scanners that examine, sort, and categorize data to detect potential violations of our Usage Policy in both user inputs and AI outputs.
- Prompt modification technology that can adjust model outputs if they might lead to harmful responses.
- A range of enforcement actions we can take in real-time if a violation is detected, including placing restrictions on accounts or removing them altogether.

Related Commitments: Thorn's Safety by Design for Generative AI; Munich Accord on Elections; Christchurch Call Commitments

Image-Based Sexual Abuse

Detection and Prevention Systems

Claude currently only produces text output and is therefore incapable of generating image-based CSAM or NCII.

We take a multi-pronged approach to detecting and preventing abusive content. For example, we may employ the following techniques:

- On our first-party services, we employ hash-matching technology to detect and report known CSAM to NCMEC that users may upload. We are implementing a similar tool for detecting NCII. Our third party partners maintain their own screening and detection systems.
- Safety classifiers on prompts and completions to identify harm.

In some instances, we will modify user inputs and Claude outputs if they are in violation of our Usage Policy.
- We are in the process of adopting interventions to avoid ingestion of CSAM, CSEM, and NCII from our training datasets.

Model Testing

We integrate policy testing that we commission from outside subject matter experts to ensure that our evaluations are robust and take into account new trends in abuse. For example, we used feedback from child safety experts at Thorn around signals often seen in child grooming to update our classifiers, enhance our Usage Policy, fine-tune our models, and incorporate these signals into testing of future models.

Monitoring and Reporting

We maintain multiple channels for identifying and reporting violative content. Our in-house Safeguards experts monitor public forums and analyze emerging abuse patterns. We have also established reporting flows that allow users to flag concerning content or model behavior.

In addition, we will continue to update the metrics above on a regular basis and will include information on child safety in our future model cards. We only publish a model card with each new model family release.

For more information, read our Progress on our Child Safety Commitments

Related Commitments: Thorn's Safety by Design for Generative AI; White House's commitments to combat Image-Based Sexual Abuse

Election Integrity

Evaluation and Testing

We take a multi-pronged approach to evaluating risks related to campaigning, lobbying, and election related misuse and abuse:

- Our Policy Vulnerability Testing (PVT) program, conducted in collaboration with external subject matter experts, examines potential election-specific risks related to misinformation, bias, and adversarial abuse.
- We also employ advanced classifiers to detect and respond to the misuse of our AI systems in elections.

Mitigations and Industry Collaboration

We collaborate with stakeholders across sectors to share threat intelligence and develop election integrity best practices. Since our models are not trained frequently enough to provide real-time election information, we've implemented several measures to ensure users can access accurate, up-to-date information.

- We’ve implemented an elections banner on Claude.ai in multiple countries to redirect users to authoritative election resources if they ask for voting information. For example, in the US we partnered with Democracy Works to direct users to authoritative election information during the relevant timeframe.
- We’ve updated Claude.ai’s system prompt to include a clear reference to its knowledge cutoff date (the date up to which Claude’s training data extends).
- While computer use is not sufficiently advanced or capable of operating at a scale that would present heightened risks relative to existing capabilities, prior to the U.S. election we put in place measures to monitor when Claude is asked to engage in election-related activity, as well as systems for nudging Claude away from activities like generating and posting content on social media, registering web domains, or interacting with government

websites.

To help others improve their own election integrity efforts and drive better safety outcomes across the industry, we released some of the automated evaluations and released multiple blog posts outlining our approach to election integrity.

Related Commitments: Munich Accord on Elections

Terrorist and Violent Extremist Content

Risk Assessment and Mitigation

We conduct pre-launch assessments and rigorous testing, informed by staff expertise and civil society, to mitigate extremist content risks. Our specialized evaluation sets are continuously updated with external insights to address evolving threats.

Transparency and Collaboration

We engage with external experts on combating extremist content through safety briefings, usage standards consultation, and policy vulnerability testing. We are partnering with the Global Project Against Hate & Extremism and the Polarization and Extremism Research Lab at American University to validate model performance on extremism and will continue to invest in similar partnerships. We will comply with requests for data in response to valid legal requests (e.g. a subpoena or a warrant).

Related Commitments: Christchurch Call Commitments

List of Voluntary Commitments

Below, you'll find a summary of our voluntary commitments. While we've highlighted the core aims of each commitment, we encourage you to review the complete commitment documents (linked) to fully understand their scope and context.

01 G7 Hiroshima Process International Code of Conduct

02 AI Seoul Summit's Frontier AI Safety Commitments

03 Seoul AI Business Pledge

04 White House's Voluntary Commitments for Safe, Secure, and Trustworthy AI

05 Munich AI Elections Accord

06 Thorn’s Safety by Design for Generative AI: Preventing Child Sexual Abuse

07 White House’s Image-Based Sexual Abuse Commitments

08 Christchurch Call Commitments

← Previous
Platform Security