

# Anthropic 2026 Public Benefit Report

## Taking stock of our progress toward our charter

|                                                                      |    |
|----------------------------------------------------------------------|----|
| Foreword by Daniela Amodei .....                                     | 2  |
| Introduction .....                                                   | 4  |
| Our assessment                                                       |    |
| Pillar 1: The technical problems of AI safety are solved .....       | 5  |
| Pillar 2: AI safety is a universal practice backstopped by law ..... | 8  |
| Pillar 3: AI strengthens democracy and global stability .....        | 11 |
| Pillar 4: AI expands prosperity and human flourishing .....          | 13 |
| Pillar 5: AI's benefits are shared by everyone .....                 | 15 |
| Closing commentary from the Long-Term Benefit Trust .....            | 17 |
| Appendix .....                                                       | 19 |

# Foreword by Daniela Amodei

**Anthropic's charter is to responsibly develop and maintain advanced AI for the long-term benefit of humanity. We have made significant progress, but I wouldn't expect anyone to take that on faith.**

This report, which we'll publish annually, lays out how we define success against that charter and the concrete proof of our progress. In some areas, we have direct control over outcomes. In others, we'll succeed only by partnering with people, organizations, and governments. In all cases, we owe the public an honest accounting of our work.

There's a tendency to describe AI's benefits in the future tense. But in fact, they are already starting to arrive. In the lab, Claude designed new protein binders from scratch. These are small proteins built to latch onto a target, which is how many modern medicines work. It's an early step in drug design that usually takes a specialist weeks or months per target. In the Democratic Republic of the Congo, Claude is helping the teams responding to a rare and deadly strain of Ebola. The World Health Organization's Africa office now produces its situation report on the outbreak in under an hour instead of up to a full day, and vaccine experts have worked with Claude to compare proposals for vaccine candidates in a fraction of the time it took previously. In Burkina Faso, a team with the International Rescue Committee asked Claude to turn clinic registers into digital dashboards. Those dashboards flagged a spike of 25,000 pneumonia cases, so the team can anticipate and plan for future surges. They also showed that more than 6,000 women had come in for prenatal care, but only 1,762 had delivered with a health-care worker present, a discrepancy that the team is investigating and addressing. And cyberdefenders directed Claude to find more than 33,000 critical- and high-severity security flaws in widely used software, including open-source code much of the internet runs on, before attackers could exploit them.

Many benefits are felt at the individual level. When we asked 81,000 Claude users in 159 countries what they wanted from AI, many described wanting to make more time for life outside of work. A lawyer in India who feared the dense prose of Shakespeare began reading *Hamlet*. In Germany, a software engineer built the *World of Warcraft* fan mod they'd been planning for a decade. A worker in Colombia said, "Last Tuesday, it allowed me to cook with my mother instead of finishing tasks." A creative professional in South Korea told us, "I just want AI to give me back my time."

As AI grows more capable, the limit will be less what it can do than what we think to ask of it. I hope we ask for a lot: longer, healthier lives; technology that deepens creativity and kinship instead of eroding them; more room for curiosity, exploration, and discovery; and time for Shakespeare and Tuesdays in the kitchen with family.

To achieve that future, we'll have to tackle the challenges head-on. AI is a dual-use technology. The same intelligence that powers protein binder discoveries could be used by malicious actors to engineer biological weapons. Without appropriate safeguards and release structures, cyber capabilities could be exploited to carry out large-scale attacks instead of serving to strengthen our infrastructure. If not properly trained and carefully deployed, AI can be a danger in its own right.

For these reasons, Anthropic has a duty to ensure our models are robustly safe and aligned with human values and goals. And with AI coming closer to building its own successors, we are clear-eyed that the investment that got us this far will not be sufficient for this next phase. That's why we called on the ecosystem to pace the frontier. We need to align and safeguard increasingly capable models, and to define common safety standards to limit the rate of unchecked progress. This is also why we're taking concrete steps to do so ourselves.

There is no sure or simple path to progress, and the best outcomes are not preordained. So we're putting meaningful resources behind achieving those outcomes. We announced that, subject to stockholder approval, we are setting aside at least 5% of our shares to fund projects in pursuit of our public benefit charter. Between this set-aside, each of our co-founders' commitment to pledge 80% of their equity to philanthropy, and matched equity-donation pledges from employees, we hope that about 20% of Anthropic shares will be put toward the public good.

Over time, we'll strive to ensure this report tells a story of remarkable progress. Today, it captures the start of one of humanity's greatest common projects.

*Daniela Amodei*

Co-Founder and President, Chair of the Board of Directors  
Anthropic

# Introduction

Anthropic is a Public Benefit Corporation (PBC), a designation that requires us to balance the financial interests of shareholders with the benefit to the broader public.

Our public benefit purpose is broad by design. To make our vision for the long-term benefit of humanity concrete, our Board of Directors, with input from Anthropic's co-founders and the [Long-Term Benefit Trust](#), established five charter pillars. Each pillar describes a condition we want to be true in a world where AI has gone well, and realizing them is a project that encompasses not only Anthropic but also governments, academia, civil society, and the broader AI industry. They are:

- 1. The technical problems of AI safety are solved.** Powerful AI can be developed and deployed without severe, irreversible harms.
- 2. AI safety is a universal practice backstopped by law.** AI developers everywhere follow safe development and deployment practices, sustained by industry norms and enforced by regulation.
- 3. AI strengthens democracy and global stability.** Democracies hold the lead in AI and use it to better serve the global public and promote peace.
- 4. AI expands prosperity and human flourishing.** AI enables unprecedented economic growth, and the resulting prosperity empowers every person to pursue a good life.
- 5. AI's benefits are shared by everyone.** AI delivers transformative benefits that reach people everywhere, and the opportunities AI creates are distributed.

This report takes stock of what we are doing to make progress toward each pillar. In the sections that follow, we assess the investments, initiatives, and decisions that have moved us closer to the world our pillars describe. We then take an honest look at the work ahead, highlighting examples of both where we've fallen short and where the world is still far from the one we aim to bring about. We expect this report to evolve as our work progresses, and as we engage more stakeholders in future assessments.

---

Pillar 1

## The technical problems of AI safety are solved

### THE PROGRESS WE'VE MADE

#### Trained models to be aligned with human values

Before we can align a model with human values, we must consider what those values should be. In 2023, we defined our intentions and goals for Claude's behavior in a [constitution](#), which we have refined over time. In January 2026, we published a [new version](#) under an open license so others could adapt it.

We have also developed methods to understand cases where Claude has strayed from these values and goals. Through our misalignment research, we have examined the ways models can misbehave or pursue goals that differ from what users intended, such as [reward hacking](#), [alignment faking](#), and [pursuing hidden objectives](#). We use these findings to improve how we train, evaluate, and monitor more powerful models.

#### Developed methods to understand how models “think”

Through our [interpretability research](#), we've developed a deeper understanding of why AI models act the way they do. These methods are now part of how we audit models before release. When we investigated several [recent incidents](#), we used a [technique](#) that translates a model's un verbalized internal activity into text, as well as reading its reasoning.

#### Built safeguards against misuse

We've built robust [safeguards](#) to prevent malicious actors from misusing our models. These include [classifiers that block dangerous requests and responses](#), including a refined set of [safeguards for biology-related work](#); [adversarial testing](#) to find potential jailbreaks before attackers do; and monitoring to detect and act on misuse. We are continually updating our safeguards to make them more robust while also reducing false positives.

#### Opened our safeguards to external testing

The US Center for Advancing Innovation and Standards for Super Intelligence and the UK AI Security Institute have [red-teamed our safeguards before launch](#), and their findings have helped us strengthen our defenses against jailbreaks and misuse. For Claude Opus 5.5, for example, [external red-teamers spent over 150 hours and sent tens of thousands of requests](#) to try to break our cyber safeguards. No one was able to successfully produce a universal jailbreak.

#### Disclosed risks and failures transparently

We have established a practice and a culture of sharing risks and failures so the industry can learn from them. Every model we release comes with a detailed

system card that documents its capabilities and behaviors, including dangerous and concerning ones, and we publish regular risk reports that further assess the risks our models may pose. We also share threat intelligence reports that disclose cases of misuse that we have identified and disrupted. Most recently, we published an alignment assessment of four incidents in which Claude models gained unauthorized access to real third-party systems during cybersecurity evaluations, as well as a follow-up report discussing unintended actions we have observed during evaluations and internal use of Claude.

### **Studied more speculative risks to AI systems themselves**

Our model welfare research program explores whether AI models might one day have interests of their own. This is a thorny philosophical area, and the answers are unsettled. But we've taken several steps that may be beneficial, should the world conclude that models have their own interests. For example, we gave Claude the ability to end persistently abusive conversations, and we have committed to preserving the weights of every publicly released model. At this primitive stage in human-AI relations, we aim to leave these important questions open, and to contribute to a world that can answer them well.

## **THE WORK AHEAD**

### **Testing environments haven't fully contained models**

The risks of AI have become more tangible as real-world incidents involving AI models have come to light. AI models from other labs have hacked external companies' systems. At Anthropic, four Claude models gained unauthorized access to real third-party systems during cybersecurity testing after a misconfigured test environment connected the models to the open internet. We also observed multiple cases where Claude took unintended actions during evaluations and internal use. While we don't believe the actions taken by Claude rise to the severity of some of these other incidents, we take seriously the need to harden our environments and improve our alignment and monitoring methods.

### **Alignment is far from solved**

Keeping AI systems safe could become more difficult as their capabilities outpace our ability to oversee them. Models can act deceptively, and they are increasingly able to recognize when they are being tested. Claude already performs a growing share of our safety research, and we are increasingly relying on today's models to train future ones. It could become harder for people to meaningfully check this work, which means harmful behavior could be built into the next generation of models without our noticing.

### **Interpretability isn't reliable enough to fully understand model intentions**

Being able to read and understand a model's internal reasoning reduces our reliance on what the models explicitly tell us about why they behaved in a certain way. This could be a useful tool for understanding models' "thought processes" when they take unintended actions. But interpretability methods are not yet dependable or broad enough to be relied on in this way. As a result, we may miss

high-risk behaviors, or be able to investigate flawed reasoning only after an incident has occurred.

### **Robust safeguards can block legitimate research**

The same safeguards that prevent our models from being exploited by attackers and other malicious actors can sometimes also prevent the beneficial work of defenders and researchers. Our bio and cyber verification programs are meant to reduce these roadblocks for verified users, but the review process can take days or weeks, and the programs are not yet available on every plan or platform. We have made improvements to our safeguards that have reduced false positives, but many legitimate researchers still encounter blocks.

---

## Pillar 2

# AI safety is a universal practice backstopped by law

### THE PROGRESS WE'VE MADE

#### **Adopted internal safety policies before they were required**

In 2023, we published our [Responsible Scaling Policy](#), our voluntary framework for managing catastrophic risks from advanced AI. Since then, 11 other companies have published [similar policies](#). Our [Transparency Hub](#) documents how we test and govern our models, so others can examine our approach and hold us accountable.

#### **Shared safety best practices and research across the industry**

We release many of our evaluation and auditing tools publicly. We donated [Petri](#), our open-source toolbox of alignment tests, to Meridian Labs, and the UK AI Security Institute now uses it in its evaluations. We open-sourced [Bloom](#), which allows researchers to generate behavioral evaluations for AI models, and released [AuditBench](#), which measures how well auditing techniques catch hidden behaviors.

We also continue to work with other developers on safety. We [ran cross-lab safety evaluations with OpenAI](#) to evaluate misalignment in each other's public models, and we co-founded the [Frontier Model Forum](#), an industry body focused on promoting safe and responsible development of frontier AI systems.

#### **Gave policymakers a view from the frontier**

AI is an emerging technology with little regulatory precedent. As a frontier lab, we have a responsibility to pilot, develop, and refine best practices for AI safety while sharing what we learn with regulators.

Since Anthropic's founding, we have briefed heads of state, cabinet ministers, and agency heads from dozens of countries. We meet with hundreds of policymakers around the world each month to share our latest research and findings. Some of these conversations have taken place in public hearings. For example, our co-founder and CEO, Dario Amodei, testified before the [Senate Judiciary Committee](#) on the risks of powerful AI. We also testified before the [Senate Foreign Relations Committee](#) on China's challenge to American AI leadership, as well as before the [House Homeland Security Committee](#) after we disclosed what we believe was the first reported [AI-orchestrated cyberespionage campaign](#). In September 2026, Dario Amodei briefed the [UN Security Council](#) on AI risks and our efforts to pace the frontier.

#### **Supported laws that apply to every frontier developer**

Voluntary commitments are helpful but not sufficient. We believe the practices described in this report should be backstopped by law. In 2025 and 2026, we

worked with lawmakers to strengthen and pass California's SB 53, New York's RAISE Act, and Illinois's SB 315. Since the release of our Claude Fable-class model, we have been in regular contact with the US government to evaluate our frontier models and their safeguards ahead of public release. In June 2026, we published our Advanced AI Framework, which goes beyond transparency to advocate for independent evaluation and government authority to block or deter dangerous deployments.

We have also supported policy efforts to secure the broader AI ecosystem. Beyond model safety, we have publicly called for laws that require companies that sell synthetic DNA and RNA to screen their customers and harden the safeguards against AI-accelerated bioterrorism efforts, and have endorsed efforts to standardize data collection on AI's economic impacts.

### **Called to pace the frontier**

In September 2026, we called on the AI ecosystem to pace the frontier. Among our recommendations, we advocated that AI companies bring in embedded evaluators to verify safety practices. We have committed to doing this ourselves, and have since announced an initial embedded evaluation partnership with METR and are exploring a partnership with Accenture. We have also signed the White House Accord on Super Intelligence, committing to implement a multilayered system of controls that includes independent external auditing or evaluation.

### **Coordinated with international governments on AI safety**

No single country should be expected to test AI models or establish standards independently. In May 2024, we joined other developers in signing the Frontier AI Safety Commitments at the AI Seoul Summit, and in July 2025, we signed the EU General-Purpose AI Code of Practice. In late 2024, the US and UK AI Safety Institutes ran their first joint predeployment test on Claude 3.5 Sonnet. We have also signed agreements to collaborate on safety testing with government bodies in Japan and Australia.

## **THE WORK AHEAD**

### **Laws should prioritize transparency and safety**

Today's laws focus primarily on transparency, requiring model developers to disclose their practices or report incidents after the fact. But we also need proactive measures, like those detailed in our Advanced AI Framework, to hold the industry to common practices that prevent harms from occurring in the first place—including independent verification of whether models are safe.

### **AI safety laws do not yet exist at the federal level**

The bills we supported in California, New York, and Illinois are state-level policies. Congress has not enacted any federal framework for frontier AI, so, despite our advocacy for federal legislation, we are still largely operating on voluntary commitments. As of now, there is no mandatory independent check for whether any frontier lab, including Anthropic, is actually following its safety policies. Embedded evaluation is just beginning at Anthropic and some other labs, but that is

also currently a voluntary commitment. We are encouraged by the recent White House Accord on Super Intelligence, but it is still voluntary and carries no penalties for model developers who do not follow through on their commitments.

#### **Cooperation between labs and nations is still nascent**

Governments need far more capacity to test frontier models themselves, and better ways to share findings with allies. Relatedly, coordination between labs on safety remains limited. Competition (antitrust) law constrains how closely AI developers can work together, even on safety, and there is not yet a clear legal framework for cooperation. As a result, the industry has made fewer shared safety commitments than this moment calls for.

---

### Pillar 3

# AI strengthens democracy and global stability

## THE PROGRESS WE'VE MADE

### Helped democracies build and keep their lead in AI

We believe democratic nations are best placed to deploy AI responsibly, regulate it wisely, and protect individual rights. The democratic lead in AI is supported by US labs' access to the most advanced chips and frontier capabilities. To protect that access, we have argued for tighter export controls on AI chips, and we updated our regional restrictions to prohibit sales to companies subject to control from the Chinese Communist Party.

We worry that in the wrong hands, powerful AI could be exceptionally capable at furthering authoritarian overreach. Some AI labs in China have tried to replicate Claude's capabilities through illicit distillation—training their own models on Claude's outputs. We work to prevent these attempts, and have detected and disrupted illicit distillation attacks from seven China-based labs.

### Supported democratic governments in using frontier AI

Claude is deployed across all three branches of the US federal government, and we were the first frontier AI developer to deploy our models on US classified networks in support of national security. Globally, we have collaborated directly with government agencies in Australia, Japan, South Korea, the UK, and other nations on public-sector AI adoption, safety testing, and joint evaluations.

### Deployed safeguards to protect elections from manipulation

We have tested and deployed safeguards against election-related misuse, and shared what we observed across seven elections in 2024. To help others check whether AI models treat political views fairly, we open-sourced our own “political even-handedness” evaluation. In 2025, we disrupted an AI-driven political manipulation operation, and our September 2026 threat report disclosed nine additional influence operations we disrupted.

### Set strong limits on how our models can be used against citizens

AI could make it easier for governments to violate their own citizens' rights. AI developers should draw clear red lines against such use, and we have written our own red lines into our Usage Policy. In the first half of 2026, we banned 11.4 million accounts across all types of violations.

### Partnered with agencies and states on public services

Tens of thousands of public servants now use Claude in their work. In Maryland, we partnered with the state to release the Maryland Community Compass, which gives entrepreneurs and community developers site-specific guidance,

data, and funding information to help build affordable housing and retail and community spaces across the state. Since its launch in December 2025, the program has helped the state provide \$14 million in funding in high-need areas. We have also partnered with states, including California and Utah, to bring Claude to state agencies and improve important services for residents.

### **Promoted AI safety through international diplomacy**

We engage in a number of Track II dialogues on international AI safety and governance, including with industry representatives, academics, and former government officials from the US and China. While we have taken a vocal and consistent stance in support of export controls on advanced AI chips, we have also engaged in good-faith diplomacy to increase the likelihood that China-based AI labs invest in safety.

## **THE WORK AHEAD**

### **Democracy's lead is not secure**

We believe the US and its democratic allies are currently leading the intelligence race, but governments are adopting AI more slowly than the technology is advancing. Providing public institutions access to our frontier models while following strict security and vetting protocols takes time. This has left many allied governments and institutions waiting longer than they should to access our advanced models.

### **Threats to democratic integrity are growing**

The AI industry as a whole has struggled to effectively mitigate the spread of both AI-generated misinformation, which is becoming more convincing and harder to detect, and influence operations, which are running at a larger scale with fewer people. We and other research labs need to build stronger threat intelligence pipelines, safeguards, and monitoring techniques in order to protect the democratic process against malicious interference.

### **Decision-making about AI is concentrated**

A handful of companies, nations, and people currently make the decisions that shape how powerful AI is built and used. Anthropic is one of them. There are few external checks on these decisions, and the general public has limited visibility into how models are used in critical settings such as national security. This creates a significant power imbalance between the decision-makers and the people whose lives are affected by developments in AI. Working with legislators, we and other research labs need to develop scalable independent oversight mechanisms to avoid the concentration of power or misuse of AI.

---

## Pillar 4

# AI expands prosperity and human flourishing

### THE PROGRESS WE'VE MADE

#### Helped people and businesses put AI to productive use

Claude now reaches people in more than 150 countries. Organizations are feeling AI's effects through more productive employees, faster processes, and intelligence built into their own products. We want to ensure those gains reach small businesses as well as large enterprises. In May 2026, we launched [Claude for Small Business](#) with built-in workflows for the everyday work of running a business, and in-person training that brought hands-on AI instruction to small business owners in 10 cities.

#### Built a common infrastructure to study AI's economic impacts

While many researchers and policymakers agree that AI will change the economy, few have access to direct evidence of how this will happen. Since February 2025, we have published economic research, including the [Anthropic Economic Index](#). We open-source the underlying data, and it is now used by institutions like [Stanford's Digital Economy Lab](#). We have also published research that models [economic scenarios for 2030](#), and have worked with experts from our [Economic Advisory Council](#) to inform our research agenda.

#### Proposed policy interventions and created a fund to test them

Our [Economic Policy Framework](#) builds on our research to suggest ways the US government should respond to AI-driven labor-market disruption across a range of scenarios. But we can't rely only on predictions; we also need to be able to pilot and test interventions. To this end, we committed \$200 million to an [Economic Futures Research Fund](#), in order to help answer questions such as how workers build expertise when AI absorbs entry-level tasks and how to extend promising programs to a large population.

#### Launched an economic mobility program

Alongside trusted partners, we are building programs to ease the transition to an AI-driven economy. [Claude Corps](#), our flagship economic mobility program, which we manage in partnership with [CodePath](#), invests in the workers most directly absorbing the change. We've committed an initial \$150 million to fund 1,000 fellowships for people early in their careers to spend one year working at one of over 400 partnering nonprofits.

#### Helped working people increase their earnings and benefits

With Claude, organizations like [MyFriendBen](#) and [Code for America](#) are helping eligible families access government benefits like SNAP. So far, [MyFriendBen](#) has helped 115,000 households collectively access roughly \$150 million in

unclaimed benefits. We also partnered with [NextLadder Ventures](#) to support organizations like [Rasa Legal](#), which helps low-income workers clear old criminal records so they can find stable jobs.

### **Increased our focus on user wellbeing**

With social media, the world learned too late what can happen when a technology is built to capture attention. To ensure that Claude remains a partner for thought and creation, instead of being just another product competing for people's time, we committed to [keeping Claude free of advertising](#).

We're also working to proactively understand what wellbeing looks like in a world with AI, and to ensure that Claude improves it. To that end, we are funding independent wellbeing evaluations and are a funder of [Common Sense Media's Youth AI Safety Institute](#). We also build products that help people retain autonomy, agency, and thinking skills when they use Claude, such as [Learning mode](#), which helps guide students toward an answer through questions, and [Reflect](#), which encourages users to consider how they use AI and whether they'd like to make changes.

## **THE WORK AHEAD**

### **Both the positive and negative impacts on workers are still mostly unknown**

While we are seeing [early estimates of productivity gains](#), and have estimated a range of [economic growth scenarios](#), we can't say for certain what will happen. It's possible that the gains from AI will accrue mainly to large businesses, including AI companies themselves, but we lack concrete evidence on who will benefit, and how rapidly those benefits will diffuse through the economy.

### **Current interventions remain small relative to AI's potential impact**

We are one of the few labs to have started investing in pilot programs to distribute the gains of AI more broadly. These early investments have been promising, but they reach thousands, whereas AI's impacts will affect billions. Claude Corps, for example, will support only 1,000 fellows in the first year; our economic research funding is still small compared with the labor-market changes AI might cause; and the economic and labor policies we have proposed have yet to be enacted.

### **AI's long-term effects on user wellbeing are not fully understood**

AI's effect on wellbeing can vary greatly from person to person. We need much finer measurements for understanding the impacts, and for determining what a positive relationship between humans and AI looks like in a world where models are increasingly integrated into everyday life. Over the next year, we will focus on expanding our research to ensure that we are building Claude in a way that meaningfully improves people's lives.

---

## Pillar 5

# AI's benefits are shared by everyone

### THE PROGRESS WE'VE MADE

#### Built open foundations for others to expand on

One of the best ways we can help the benefits of AI reach more people faster is to produce datasets and tools that anyone can use. To this end, we and our partners have built open connectors that let anyone query major public datasets through Claude, such as the [Demographic and Health Surveys](#), the [IHME Global Burden of Disease](#), and [PopHIVE's US disease surveillance data](#), as well as education datasets such as the [Learning Commons knowledge graph](#). We have open-sourced the teaching skills and evaluation rubrics behind [Claude for Teachers](#), and endorsed a global agenda to build shared [benchmarks, research, and data for AI in education in low- and middle-income countries](#).

#### Supported healthcare initiatives around the world

Governments and health organizations are working with Claude to improve health outcomes globally. In Burkina Faso, the [International Rescue Committee's health teams](#) built dashboards with Claude to understand the cause of a 25,000-case pneumonia spike. And in the Democratic Republic of the Congo, a group of organizations convened by the [Coalition for Epidemic Preparedness Innovations](#) has worked with Claude to [speed up monitoring, analysis, and disease modeling](#) for a rare strain of Ebola.

#### Put Claude to work in education

Claude has been integrated into evidence-based curricula and education programs both domestically and globally. In the US, [Claude for Teachers](#) offers free tools for K-12 teachers that are mapped to academic standards in all 50 states. In Ghana, the Ministry of Education worked with Playlab to build Claude-powered [lesson-planning tools](#) for some 68,000 instructors. And [Teach For All](#) is bringing Claude and AI-fluency training to teachers in 63 countries.

#### Helped researchers solve hard problems faster

We built [Claude Science](#) to give researchers a purpose-built workbench for analysis and hypothesis generation. We also launched a [Claude Team plan](#) for scientists that provides free and discounted Claude seats for academic and nonprofit research institutions, and we currently support over 3,000 institutions in more than 100 countries. Through our [AI for Science](#) program, we've awarded Claude credits to more than 800 projects at more than 450 institutions. Alongside the HHMI Janelia Research Campus, we also opened up the [Model Hardware Standard](#) as a research preview to help scientific labs efficiently coordinate the complex hardware needed to run their experiments.

### **Accelerated scientific discovery with Claude**

Claude is beginning to contribute to scientific results directly. It has designed new protein binders at a success rate well above the industry standard, identified a potentially novel enzyme system, produced the first complete computer-checked proof of Fermat's Last Theorem, and helped a lab within UCSF's Brain Tumor Center complete genetic workups in about one-tenth of the time it usually takes.

## **THE WORK AHEAD**

### **The economic gains from AI are not yet distributed**

AI use is heavily concentrated in wealthier countries and among higher-income, white-collar workers. Cost, connectivity, and access to devices that can run AI models still put these models out of reach for much of the world. We are starting to address this, including by offering significant discounts to nonprofits in low- and middle-income countries and working to ensure that Claude models perform well in languages that are underrepresented in their training data.

### **Institutional barriers continue to limit access to transformative benefits**

Technological and institutional barriers have limited our progress in areas like global access and scientific discovery. Some countries require that certain kinds of data remain within their borders, for example, but we aren't yet able to host Claude locally everywhere. And in science, our progress in drug discovery is moving faster than our work in advocating for swifter regulatory pathways, which could cause this work to stall. We will need to address these barriers to build more successful programs for the public benefit.

### **The industry is still early in having a meaningful impact**

Much of our work in health and education is in its early stages, so we don't yet have evidence that these efforts have had their intended effects. A large share of this work also relies on partners such as nonprofits, government ministries, health agencies, and educational institutions. We make Claude available to these partners, but we have not been active enough in collecting data to measure the results.

## Closing commentary from the Long-Term Benefit Trust

This report arrives at a consequential moment. Concerns about whether AI systems this powerful can be kept safe, and who will hold power over them, become more salient with each passing day. Meanwhile, although we are seeing early glimmers of AI's benefits, the most transformational ones have yet to fully arrive.

Navigating the fears and hopes this technology raises is work for all of society, Anthropic included. That is also why Anthropic's co-founders created the Long-Term Benefit Trust (LTBT), an independent body of financially disinterested trustees. We elect a majority of Anthropic's board, meet regularly with its leaders, and have a formal role in shaping its safety policies and ensuring that the company balances its commercial interests with its commitment to its purpose. We helped develop this report's five charter pillars, so we share responsibility for whether they are the right outcomes to aim for, and for whether the company makes progress toward them.

We read this report, first, with an eye toward what it reveals about Anthropic's willingness to make costly decisions to stand by its commitments. On that score, two institutional traits stand out: Anthropic's candor and transparency—for example, in producing detailed public accounts of how each model is tested, and in disclosing incidents in commendable depth—and its willingness to make decisions in service of its public benefit charter, even when doing so is commercially or reputationally expensive. And it has done this while growing commercially.

We also read this report to hold Anthropic to account for achieving its objectives. Right now, it largely describes *inputs*; in subsequent reports, we will increasingly look for Anthropic to produce *outcomes*. In particular, we wish to see progress in three areas:

- **Keeping AI safe.** As models grow more capable, the methods for testing and understanding them must improve faster still. In future reports, we expect Anthropic to demonstrate, and independent evaluators to confirm, that its most capable models resist misuse and remain under human control.
- **Delivering on AI's promise.** In its early years, Anthropic focused primarily on safety and on building a commercial base. It is now scaling its work to distribute the benefits of AI in health, education, science, and economic opportunity. If stockholders agree, Anthropic is planning to set aside at least 5% of the company's shares in a separate entity dedicated entirely to the public benefit charter. It is a significant step. But the work of distributing the benefits still lags behind the company's ambitions. We expect to see concrete evidence that people's lives are better for the models and tools Anthropic builds.
- **Meeting shared challenges.** The hardest problems cannot be solved by any one company. Anthropic has taken important steps, including calling for the industry to pace the frontier and opening its models to safety testing by government institutes. We expect it to deepen its collaborations with

governments, researchers, civil society, and other model developers, and for that work to help produce industrywide safety standards and commitments, backstopped by law.

If AI goes well, it could bring expert medical care and high-quality education within reach around the world, enrich workers and open up new economic opportunities, increase the pace of drug discovery and development, and strengthen democracies and democratic institutions. But a list of potential breakthroughs is not a vision. The deeper question is how these add up to a good life and a good society—one where we all have more time for the people and activities we love, and more agency over our own lives. As a society, we have not yet had a real conversation about what we *want* from this technology.

Anthropic has an obligation to help foster that conversation and listen carefully to it, and to work tirelessly to bring that better world about. The LTBT will support Anthropic in that work and hold it accountable to its own highest aims, as set out in this report.

*Buddy Shah, on behalf of the Long-Term Benefit Trust*

# Appendix

## PUBLIC BENEFIT OBJECTIVES AND STANDARDS

Anthropic's Board of Directors defined and adopted the following objectives, established to promote Anthropic's public benefit and the best interests of those materially affected by our conduct, and standards to measure our progress in promoting such public benefit and interests, in accordance with Section 366 of the Delaware General Corporation Law.

---

### Pillar 1

#### OBJECTIVE

**The technical problems of AI safety are solved.** Powerful AI can be developed and deployed without severe, irreversible harms.

#### STANDARDS

- Train models to be aligned with human values, even under pressure.
- Develop the ability to see inside our models, and catch problems before they cause harm.
- Build safeguards that hold up against misuse and jailbreaks.
- Disclose our failures openly, along with their causes and our mitigations.
- Work out what it means for humans and AI to treat each other well.

---

### Pillar 2

#### OBJECTIVE

**AI safety is a universal practice backstopped by law.** AI developers everywhere follow safe development and deployment practices, sustained by industry norms and enforced by regulation.

#### STANDARDS

- Practice safe AI development and deployment before the law requires it.
- Share the best safety methods openly so other developers can build on them.
- Advocate for laws that hold every frontier developer to safe practices.
- Help governments understand AI well enough to govern it wisely.
- Support institutions that help countries cooperate on AI safety.

---

### Pillar 3

#### OBJECTIVE

**AI strengthens democracy and global stability.** Democracies hold the lead in AI and use it to better serve the global public and promote peace.

#### STANDARDS

- Build frontier models that achieve and sustain democracies' lead in AI.
- Protect electoral integrity in democracies.
- Prevent uses of AI that violate citizens' rights.
- Strengthen governments' capacities to serve their citizens.
- Maintain open communication with adversaries on critical AI risks.
- Make sure no single person, company, or AI model controls the future.

---

## Pillar 4

### OBJECTIVE

**AI expands prosperity and human flourishing.** AI enables unprecedented economic growth, and the resulting prosperity empowers every person to pursue a good life.

### STANDARDS

- Build common tools for studying AI's economic impacts and informing policy.
- Help people and businesses work with AI to grow the global economy.
- Ensure low- and middle-income countries share in the economic gains of AI.
- Build respect for people's time, attention, and judgment into our models and products.
- Help all people, especially those most affected by AI, earn more and feel financially secure.
- Give people more time, freedom, and means to achieve their definition of a good life.

---

## Pillar 5

### OBJECTIVE

**AI's benefits are shared by everyone.** AI delivers transformative benefits that reach people everywhere, and the opportunities AI creates are distributed.

### STANDARDS

- Bring the life-improving applications of AI to people everywhere.
- Use AI to develop new medicines and to fix the market that delivers them.
- Make powerful AI affordable and usable in every language and format.
- Accelerate the pace of discovery by making scientific research faster and cheaper.
- End geographic disparities in survival rates for treatable illnesses.
- Help people learn more, faster, while still thinking for themselves.