11 MARZO 2026

Introducing The Anthropic Institute

We're launching The Anthropic Institute, a new effort to confront the most significant challenges that powerful AI will pose to our societies. The Anthropic Institute will draw on research from across Anthropic to provide information that other researchers and the public can use during our transition to a world containing much more powerful AI systems.

In the five years since Anthropic began, AI progress has moved incredibly quickly. It took us two years to release our first commercial model, and just three more to develop models that can discover severe cybersecurity vulnerabilities, take on a wide range of real work, and even begin to accelerate the pace of AI development itself.

We predict that far more dramatic progress will follow in the next two years. One of our company's core convictions is that AI development is accelerating: that the improvements we make are compounding over time. Because of this, extremely powerful AI, like the kind our CEO Dario Amodei describes in *Machines of Loving Grace*, is coming far sooner than many think.

If this is right, society is shortly going to need to confront many massive challenges. How will powerful AI systems reshape our jobs and economies? What kinds of opportunities for greater societal resilience will they give us? What kinds of threats will they magnify or introduce? What are the expressed "values" of AI systems and how will society help companies determine what the appropriate values are? And, if the recursive self-improvement of AI systems does begin to occur, who in the world should be made aware, and how should these systems be governed?

The Anthropic Institute's goal is to tell the world what we're learning about these challenges as we build frontier AI systems, and to partner with external audiences to help address the risks we must confront. Whether our societies are able to do so will

determine whether or not transformative AI delivers the radical upsides that we believe are possible in science, economic development, and human agency.

The Institute is led by our co-founder Jack Clark, who will assume a new role as Anthropic's Head of Public Benefit. It has an interdisciplinary staff of machine learning engineers, economists, and social scientists, bringing together and expanding three of Anthropic's research teams: the Frontier Red Team, which stress-tests AI systems to understand the outermost limits of their current capabilities; Societal Impacts, which studies how AI is being used in the real world; and Economic Research, which tracks its impact on jobs and the larger economy. The Institute will also incubate new teams, and is currently working on efforts around forecasting AI progress and better understanding how powerful AI will interact with the legal system.

The Institute has a unique vantage point: it has access to information that only the builders of frontier AI systems possess. It will use this to its full advantage, reporting candidly about what we're learning about the shape of the technology we're making. At the same time, the Institute is a two-way street. It will engage with workers and industries facing displacement, and with the people and communities who feel the future bearing down on them but are unsure how to respond. What we learn will inform what the Institute studies, and how our company as a whole chooses to act.

The Anthropic Institute has made several founding hires:

- Matt Botvinick, a Resident Fellow at Yale Law School and previously Senior Director of Research at Google DeepMind and Professor in Neural Computation at Princeton, is joining the Institute to lead its work on AI and the rule of law.
- Anton Korinek is joining the Economic Research team, on leave from his role as Professor of Economics at the University of Virginia, to lead an effort studying how transformative AI could reshape the very nature of economic activity.
- Zoë Hitzig, who previously studied AI's social and economic impacts at OpenAI, is joining to connect our economics work to model training and development.

We're also hiring. The Anthropic Institute is building out a small analytical staff who will work to pull various parts of our research agenda together and broadcast our work to the world. You can read more here.

Expanding Anthropic's Public Policy team

Alongside launching The Anthropic Institute, we're expanding our Public Policy organization.

Public Policy focuses on the areas where Anthropic has defined priorities and perspectives, including model safety and transparency, energy ratepayer protections, infrastructure investments, export controls, and democratic leadership in AI. Sarah Heck, who joined Anthropic as our Head of External Affairs, will lead this team as Head of Public Policy. Before Anthropic, Sarah was Head of Entrepreneurship at Stripe, a financial technology firm, and previously led global entrepreneurship and public diplomacy policy at the White House National Security Council.

We're growing our Public Policy team to help inform and shape AI governance around the world. We're opening our first office in DC this spring, and are quickly expanding our global policy footprint. You can see our current openings here.