The Societal Impacts Team at Anthropic is Hiring

In this post, we describe the Societal Impacts team's **vision**, discuss the **problems we're currently thinking about**, and provide a **FAQ** for candidates considering applying. Additionally, for the curious, we leave an annotated list of our **previous research and policy contributions** at the end**.**

Vision

We are a safety research team within Anthropic whose goal is to ensure that AI interacts positively with people: from individuals to communities to our society as a whole. We are a technical research team that tackles this goal using tools from machine learning as well as from across the social sciences, human-computer interaction, policy, and other fields. Our work has taken three directions:

- **Socio-technical alignment** asks questions like: How do we choose the human values we want our models to hold? How do we make it easier for people to specify their subjective and nuanced values? How should our algorithms operate in the face of conflicting or ambiguous values? How are our models used and misused in the wild? How can we anticipate future societal risks of our systems? How do we develop precise experiments and evaluations to answer all of these questions?

- We believe providing trustworthy **policy-relevant evaluations** about topics policymakers care about will lead to better policy outcomes. To do this, we work closely with our colleagues in Public Policy and Trust & Safety to gather evidence that both Anthropic and the external policy community can use when making decisions about the future of AI. This work can be *reactive*, in response to questions from policymakers (such as our work on mitigating discrimination from our models), as well as *proactive*, attempting to anticipate needs coming down the line (such as our work on election integrity).

- We **identify and fill research gaps** in Anthropic's broader AI safety research mission by working across different research teams. For example, we heavily invested in research engineering tools to make designing and running evaluations at scale as painless as possible. We were the first team within Anthropic to establish processes and procedures for red-teaming our systems, in close collaboration with our alignment and human feedback teams. We believe in generating useful resources and open-sourcing them so the broader public can critique and build on our work.

Problems we're currently thinking about

Here's a small sample of some of the problems we are currently working on, or are excited to explore:

- **How do we help maintain election integrity?** As part of our broader efforts to prepare for global elections in 2024, we're building evaluations and testing how our models perform against misuse in election-related settings. We're also designing experiments that measure how persuasive our models are, with an eye towards how models could be misused.

- **How do we make AI systems that augment rather than replace humans?** We're exploring new ways that people and models can interact outside of the dominant paradigms. For example, we are exploring how models can elicit better and deeper preferences and values from people so that they can be more aligned with human intent.

- **How do we use mechanistic interpretability to help with AI alignment?** We're collaborating with Anthropic's Interpretability team to better understand the inner workings of our models so that we can develop new ways of anticipating and mitigating safety failures.

- **How do we maintain oversight over increasingly agentic systems?** We are collaborating with Anthropic's Frontier Red Team and Alignment Science team to explore new paradigms for how our frontier models can be used to solve increasingly complex tasks. These new paradigms come with new threat models that need to be articulated.

- **How can we bridge the gap between the research and product?** We are collaborating with Anthropic's Product teams in order to understand how people use our systems, what they value, and what this means for the future of work. Additionally, we're collaborating with our Trust & Safety team to understand how our models are misused and how to develop mitigations.

Join us!

We have many more ideas, and we need your help to make them a reality. If you're interested in joining us, apply here today!

Frequently asked questions

1. **How big is the team?**

We are currently 6 people. Esin Durmus, Alex Tamkin, Miles McCain, Kunal Handa, Saffron Huang, and Deep Gangul (hiring manager)

2. **What is the growth plan?**
   We are currently only looking to add 1 research engineer ASAP! We're looking to add up to 4 research scientists by the end of 2025.
3. **Do you work across teams?**
   Yes. We collaborate strongly with our other research teams including: Policy, Alignment, Interpretability, and Safeguards (formerly Trust and Safety). Anthropic is a "team-science" place, and the borders between teams can be porous in the best ways—this allows us to get a lot done quickly! Furthermore, we spend a lot of time framing projects, prior to execution (with a light-weight buy-in process), such that they can be executed quickly. We try to design experiments where even a *null* result is interesting!
4. **What's the difference between the societal impacts team and the alignment science team?**
   We focus on more socio-technical alignment, e.g., figuring out what values/norms/rules to align to, and building new human-centric methods for understanding and aligning our models. The Alignment teams focus more on algorithm development and mitigating catastrophic risk, according to AI Safety Levels and responsible scaling plans. Finally, the societal impacts team spends relatively more time engaging in AI policy work. Nevertheless, both teams often use the same tools, research methods, and experimental designs. We also collaborate closely!
5. **How do you pick what projects to work on?**
   We have three high level approaches to picking projects to work on: 1) A policy maker asks us a question we simply do not know the answer to, and it seems important for us to know the answer 2) Our Policy team helps us identify what kind of research policymakers might be interested in, either now or in the near future 3) We identify research areas nobody else at Anthropic is working on that are consistent with our Core Views on AI Safety. Every member on the team pitches research directions, and as a group, we decide on what to prioritize every ~6 months in a relatively light weight process.
6. **I'm interested in doing more of the economic impacts of AI.**
   So are we! Anthropic is only just starting to create a research and policy program on the economic impacts of AI, drafting off of the Societal Impact's teams foundational research that powers the Anthropic Economic Index.
7. **What kinds of research backgrounds do people on the team have?**
   At the moment, we all have some background at the intersection of computer science, social science, and natural sciences. Heavy on computer science—our work involves interacting with our quickly-moving research and production codebases. Strong programming abilities are important. This is why our on-site interviews focus much more

on your programming abilities than your research abilities. Our hiring manager screen focuses on your research skills. We all also care about AI policy and trying to effect change at the institution level.

8. **I'm a researcher and love the vision but am not ready to apply. Can we collaborate?**
   We are open to external collaboration; however, we get so many requests that we often don't feel like we can maintain more than ~1 collaboration per year. We want to be good collaborators, so we often have to say no to really exciting possible collaborations.
9. **Are you open to candidates outside of the Bay Area?**
   No. The entire team is based in San Francisco. We come into the office frequently, love in-person brainstorming, and it is imperative to keep this culture as we scale. Trying to do team research remotely just seems less effective than in person.

Previous research and policy contributions

Below are our previous papers, blogs, datasets, policy outputs, and media mentions. Each item is annotated below with a short description of how the work connects back to our vision.

- Predictability and Surprise in Large Generative Models [FAccT '22] {Quanta}
  In our very first paper, we discuss a paradoxical implication of scaling laws—namely that there are strong technical and economic incentives for making increasingly larger and more powerful systems; however, larger models may acquire capabilities or cause harms that are difficult to anticipate in advance. We discuss the policy implications of this paradox and suggest policy recommendations that we followed up on in subsequent work (see below!). This paper corresponds to the start of our team's efforts to build rigorous socio-technical evaluations and we also used these evaluations in Anthropic's first research paper, to test the efficacy of our earliest alignment techniques to train our models to be helpful, honest, and harmless.

- Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned [Arxiv, '22][Red Team Data]
  One of our policy recommendations from "Predictability and Surprise" was to invest in red teaming—or adversarially testing models for harm prior to release. In our next paper, we thoroughly described Anthropic's early efforts and challenges in building scalable and repeatable processes for red-teaming our systems with crowdworkers. We showed how red-teaming can contribute to finding, measuring, and mitigating possible harms. We also open-sourced all of our red-teaming data so other developers can use the data to make their models safer. In addition to publishing our research, we shared our findings with the policy community, drawing connections between our work and the AI Risk Management Framework developed by the National Institute of Standards and

Technology (NIST). Our work on red-teaming also allowed us to contribute our learnings in a working group that designed the AI Red Teaming event at DEFCON last summer.

- The Capacity for Moral Self-Correction in Large Language Models [Arxiv '23]{MIT Tech Review}
  Another of our policy recommendations from "Predictability and Surprise" was to invest in building quantitative evaluation suites to measure social biases. In our third paper, we did this *and* used our evaluation suite to find something that surprised us: you could *significantly* decrease social biases with simple prompting methods, but only with sufficiently large models trained with RLHF. Our work on moral-self correction helped motivate our broader effort on Constitutional AI and some Trust & Safety functionality for our product. Our work had two other technical positive externalities. First, we found the evaluations in this paper so painful to implement with our existing tech stack, that we re-built a new evaluation tech-stack that is widely used by researchers and engineers across the organization. Second, we also built Claude for Sheets, so that people without programming skills can also explore running simple evaluations in Google Sheets. This turned into a product.

- Towards Measuring the Representation of Subjective Global Opinions in Language Models [Arxiv '23][GlobalOpinionQA Dataset, Interactive Web Version]
  As Anthropic began to consider rolling out Claude to more people around the world, we wanted to understand how it might operate in non-US contexts. This led us to ask the question: how representative are Claude's responses to people's opinions on global issues across different countries? We developed an evaluation to measure this and found that Claude's responses are most similar to the opinion distributions of people from the USA, Canada, and some European and South American countries. We found that when we prompted Claude to respond to survey questions as though it were from another country, Claude was quite steerable; however, upon further investigation we found evidence that Claude was sometimes relying on cultural stereotypes in this condition. When we changed the language we prompted Claude in (since linguistic cues are powerful cultural cues) we found the opposite effect: Claude's opinions on global issues remained consistent with American opinions. In our paper, we described the myriad challenges and simplifying assumptions we made in order to design our sociotechnical evaluation, which we subsequently open-sourced for others to use and build on.

- Opportunities and Risks of LLMs for Scalable Deliberation with Polis [Arxiv, '23]
  As we began to prepare for global elections happening throughout this year, we started to think about what kind of impact AI might have on democratic processes. This is a challenging, multifaceted question, so we started small in order to make the problem tractable. We began by collaborating with the Computational Democracy Project on a narrowly scoped project to explore how Claude could be used to facilitate online digital town-halls with the Polis platform. Our findings were surprising—Claude could effectively

predict people's votes on niche political issues *and* also effectively summarize the outputs of long deliberative processes. This project was our team's first research into the capabilities of our long context models, as digital town halls generate outputs that are too large to fit into the context windows of most models available at the time. This project was also our first external research partnership, and we built up our collaboration muscle as a result.

- Collective Constitutional AI: Aligning a Language Model with Public Input [Anthropic Blog '23, FAccT '24]{New York Times}
  Working with the Computational Democracy Project gave us a crazy idea. What if instead of having Anthropic decide the ethical principles in Claude's AI constitution, we asked the public what values Claude should abide by? Could we crowd-source these values via an online deliberation process involving a representative sample of the American public? What would happen if we then A/B tested a model trained using the crowd-sourced constitution versus a model trained using the Anthropic-curated constitution? In collaboration with the Collective Intelligence Project we explored how more democratic processes might influence AI development. In our experiment, we discovered areas where people both agreed with our in-house constitution, and areas where they had different preferences. We found that when we trained a model using the crowd-sourced constitution and compared it to one trained using the Anthropic constitution, the former model was generally less biased and equally capable as the standard Anthropic model. This project surfaced a new principle related to being respectful of people with disabilities that has now been incorporated into Claude 3. Furthermore, our work paved the way for a novel empirical approach to answer the question: "whose values should we align our systems with?"

- Evaluating and Mitigating Discrimination in Language Model Decisions [Arxiv '23] {VentureBeat} {TechCrunch} [DiscrimEval Dataset]
  As the capabilities of language models advance, many people worry that they may be used to make high-stakes societal decisions about people, and might not do so in an equitable way. Policymakers are actively preparing for this possibility in their work on the recent Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence and the EU's AI Act. In an effort to make progress, we developed a new evaluation technique to measure discrimination when language models are used to make high-stakes decisions about people. With our methodology we found patterns of both positive and negative discrimination in Claude 2.0 across 70 potential applications of language models across the economy. We also expanded on our previous work on Moral Self-Correction (see above) and found simple yet effective prompting techniques to mitigate discriminatory outputs. For both this research and our work on Collective Constitutional AI, we partnered closely with the Policy team to share our findings and experiences with the policy community through a series of memos and briefings.

- Challenges in Evaluating AI Systems [Anthropic Blog '23]

Through our policy work, we found that many policy makers are developing AI governance initiatives that heavily rely on technical evaluations of AI systems. To help policymakers in this effort, we outlined challenges that we have encountered while evaluating our own models (over the course of three years) in order to give readers a sense of what developing, implementing, and interpreting model evaluations looks like in practice. We documented the challenges we faced designing evaluations from all the above papers, in addition to lessons learned from our working with third-party evaluation frameworks like BIG-bench and HELM, implementing model-written evaluations, and participating in a third-party audit with a non-profit.

- Measuring the Persuasiveness of Language Models [Anthropic Blog '24]
  We study persuasion because it is a general skill which is used widely within the world—companies try to persuade people to buy products, healthcare providers try to persuade people to make healthier lifestyle changes, and politicians try to persuade people to support their policies and vote for them. Developing ways to measure the persuasive capabilities of AI models is important because it serves as a proxy measure of how well AI models can match human skill in an important domain, and because persuasion may ultimately be tied to certain kinds of misuse, such as using AI to generate disinformation, or persuading people to take actions against their own interests. In our research, we found a clear scaling trend across model generations: each successive model generation is rated to be more persuasive than the previous. We also find that our latest and most capable model, Claude 3 Opus, produces arguments that don't statistically differ in their persuasiveness compared to arguments written by humans.