Name: Center for Human-Compatible AI
Location: UC Berkeley
Founded in: 2016
Team size: ~50
2020 Funding: ~$2M

Center for
Human-Compatible
Artificial
Intelligence

Organization website

Open Philanthropy Assessment

Who Are They?

The Center for Human-Compatible AI (CHAI) is an academic research centre at University of California, Berkeley that carries out technical and advocacy work to help ensure the safety of AI systems and build the field of future AI researchers.

What problem are they trying to solve?

In building advanced machine intelligence, we would forfeit our position as the most intelligent force on the planet, and we are currently doing so without a clear plan. Given the potential benefits we could enjoy if the transition to advanced general AI goes well, successfully navigating the transition to advanced AI seems to be one of the most important challenges we will face.

Artificial intelligence research is concerned with the design of machines capable of intelligent behaviour, i.e., behaviour likely to be successful in achieving objectives. The long-term outcome of AI research seems likely to include machines that are more capable than humans across a wide range of objectives and environments. This raises a problem of control: given that the solutions developed by such systems are intrinsically unpredictable by humans, it may occur that some such solutions result in negative and perhaps irreversible outcomes for us. CHAI's goal is to ensure that this eventuality cannot arise, by refocusing AI away from the capability to achieve arbitrary objectives and towards the ability to generate provably beneficial behaviour. Because the meaning of beneficial depends on properties of humans, this task inevitably includes elements from the social sciences in addition to AI.

What do they do?

CHAI's goal is to develop the conceptual and technical wherewithal to reorient the general thrust of AI research towards provably beneficial systems. CHAI aims to do this by developing a "new model" of AI, in which (1) the machine's objective is to help humans realise the future we prefer; (2) the machine is explicitly uncertain about those human preferences; (3) human behaviour provides evidence of human preferences. This is unlike the standard model for AI, in which the objective is assumed to be known completely and correctly.

CHAI's research spans computer science, psychology, economics and other areas, and CHAI faculty and Principal Investigators include academics at various world-leading American universities. CHAI is one of the few academic research centres that is focused solely on the safety of advanced AI. Given that the field of AI safety is controversial, especially in relation to global catastrophic risk, CHAI's ability to afford the topic academic legitimacy is likely to be especially valuable. CHAI faculty members and Principal Investigators have a very strong reputation in the field.

What would they do with more funding?

Technical AI safety research

There are a number of research projects that CHAI would like to pursue.

Graduate student or postdoc salary

With an extra $100,000, CHAI could fund two additional graduate students for one year.

Research Internships

CHAI hosts 6-8 summer interns per year. This programme has three main benefits:

- Increases CHAI research output.
- Builds the AI safety research community globally.
- Enables individuals to test their fit for AI safety work and gain experience to allow them to pivot towards it.

Usually the interns work with CHAI for three months on a stipend of $3,000 per month to cover their living and travel

Visiting researchers

CHAI offers experienced researchers the opportunity to work with CHAI for limited periods of time, such as for one year. For example, CHAI's current 'Machine Learning Research Engineer' position provides engineering expertise and allows CHAI to run Machine Learning experiments more easily, which helps speed up research output. For around $150,000, CHAI would be able to fund a one-year visiting researcher of this kind.