Artificial intelligence (AI) is probably the most important technology we ever develop. Ensuring it is secure and used beneficially is one of the best ways we can safeguard the long-term future.

AI is already incredibly powerful: it’s used to decide who receives welfare, whether a loan is approved, or whether a job applicant receives an interview (which may even be conducted by an AI).

It’s also a research tool. In 2020, an AI system called AlphaFold made a “gargantuan leap” towards solving problems in protein folding, which scientists have been working on for decades.

Despite these impressive accomplishments, AIs don’t always do what we want them to. Amazon used an AI to screen resumés, thinking this would increase the fairness and efficiency of their hiring process. Instead, they discovered the AI was biased against women. It penalised resumés containing words like "women's" and "netball," while favouring language more frequently used by men, such as "executed" and "captured." This was not intended, but that may be of little comfort to the women whose applications were rejected because of their gender.

Ensuring AI is used to benefit everyone is already a challenge, and it's critical we get it right. As AI becomes more powerful, so does its scope for affecting our economy, politics, and culture. This has the potential to be either extremely good, or extremely bad. On the one hand, AI could help us make advances in science and technology that allow us to tackle the world's most important problems. On the other hand, powerful, but out-of-control AI systems (or "misaligned AI") could result in catastrophe for humanity. Given the stakes, working towards beneficial AI is a high-priority cause that we recommend

supporting, especially if you care about safeguarding the long-term future.

# How we can promote Beneficial AI

There are two challenges we need to solve to ensure AI is beneficial for everyone:

- The *technical* challenge: how can we make sure powerful AI systems do what we want them to?
- The *political* challenge: how can we ensure the wealth created by AI is distributed fairly, and incentivise AI companies to build AI safely?

## Technical challenge: Ensuring AI systems are safe

These technical problems are a significant challenge that we could make progress on, i recommend donating to the Center for Human-Compatible Artificial Intelligence (CHAI). Their research agenda is to find a new model of AI in which the AI's objective is to satisfy human preferences. The hope is that by doing so, we can make progress on the specification problem described above. So far, CHAI has produced an extensive amount of published research, developed the field of AI safety by funding and training nearly 30 PhD students,

## Political challenge: Promoting beneficial governance

If AI development only aims to make a profit, there is a risk we will see what Oxford Professor Allan Dafoe calls *value erosion*, where companies are incentivised to progress quickly, rather than safely. This is because they would capture a significant proportion of the rewards of powerful, safe AI systems — whereas if the AI is unsafe, they would only be one of many who pay the price.

Even if AI is safe in the technical sense — meaning it is performing as intended — we still need to make sure it is used to benefit everyone. There is some risk it would not be. Poorly governed but powerful AI systems could result in unprecedented wealth inequality, or lock in the (potentially undesirable) values of a handful of people, with no consultation from the public.

These problems may seem abstract and mainly focused on hypothetical issues of AI systems with capabilities far beyond those that exist today. But as we saw with the gender bias in Amazon's algorithm, ensuring that today's AI is used beneficially is already a major challenge.

# Reference:


1. https://www.openphilanthropy.org/blog/some-background-our-views-regarding-advanced-artificial-intelligence#Sec1

2. For a good, non-technical introduction to these concrete problems, i recommend watching Robert Miles on Computerphile ( https://youtu.be/AjyM-f8rDpg )

3. I recommend reading Nick Bostrom's *Superintelligence*, which covers these scenarios in detail ( https://www.google.com.au/books/edition/Superintelligence/7_H8AwAAQBAJ?hl=en&gbpv=1&printsec=frontcover )

4. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G