List of research proposals Otto Barten FHI

AI Policy

Delayed Singularity: an overview of arguments for and against superintelligence postponement

In his recent book The Precipice, Toby Ord shortly discusses the possibility of postponing superintelligence, until the risks are handled better. Would this be a feasible option? Which policies and coordination mechanisms may be useful in achieving this? And do the benefits, such as possibly better risk handling and increased wisdom, outweigh the costs?

Superintelligence skepticism

Although an academic consensus on the possibility of superintelligence and some of its consequences slowly emerges, most non-academics are still skeptical. This means there is a large gap between the most likely future as the research community sees it, and the future most of humanity is preparing for. This research will investigate superintelligence skepticism and seek to find answers to questions such as:

- How widespread is it among different groups? E.g. by profession, country of residence, age, sex, income, etc.
- In what way is it changing over time? Can we quantitatively model expected superintelligence skepticism over time? Can such a model forecast how large skepticism

will be at important points in the development of AGI? This may be used to assess expected behaviour of e.g. democratic decision processes. Climate change skepticism could serve as an analogous situation.
- What is the effect of different kinds of information on skepticism?
- What are the effects of skepticism, for example on policy? Both negative and positive effects will be investigated.

What can we learn from how democracies work, for AGI alignment?

If we can build safe AGI, it will need input from human beings. Currently, the most established input channel for country-level desires in many countries is the democratic decision making process. Would it be possible to integrate AGI utility function generation with outcomes of this process? What would be challenges and opportunities?

Superintelligence

Easy goals: controlling superintelligence using non-optimizing utility functions

The paperclip optimizer could partially be dangerous because it is an optimizer. If its utility function would not be to create as many paperclips as possible, but instead to create 100 paperclips, it would be much harder to see how that could be a danger for humanity. Building on this idea: would it be possible to explore superintelligence step by step, by creating achievable goals which require a certain well defined amount of intelligence only mildly above a human being? Or could existential risk even be reduced altogether by setting achievability standards for goals as a development best practice?

Unbasic AI drives: when does evolutionary pressure stop?

Omohundro has in The Basic AI Drives proposed that evolutionary pressure makes any AI design strive for certain outcomes. This may be true in cases where there is evolutionary pressure on AI. However, if there is a single AGI which, as a relatively easy first step, bans all other AGIs from development, it could be argued that this evolutionary pressure stops there. Will Omohundru's basic AI drives still hold in such a situation?

Personal strategies in the AGI century

Many people are planning fairly long term: they buy houses with mortgages lasting decades, save up for their pensions, write wills, and raise children. According to academic consensus, AGI development could well take place within these time spans. As an individual, what would be an optimal strategy for one's own long term future, given that this risk, but also opportunity,

exists? It could be a conclusion that it seems a good idea to do something with this situation as an individual. In that case, results from this research may motivate the public to act on the best academic knowledge, which could in turn be helpful (or harmful) for existential risk.

Risk quantification

Quantifying AGI risk

In Toby Ord's recent book The Precipice, existential risk because of unaligned AGI is estimated at 1/10 for this century. This number, which is important for his main arguments, is however hardly substantiated. There is some research on which actions would need to be taken before AGI can become an existential threat. The aim of this project is to put probability numbers on these steps, thereby substantiating the important AGI existential risk estimate, and making it more open to fruitful discussion.

What would be residual existential risks in case aligned AGI is developed technically?

According to the orthogonality thesis as expressed by Nick Bostrom's Superintelligence, goals and capacities are separate dimensions. If we manage to generate safe AGI capacity, could we still suffer existential risk from badly formulated goals? Can we estimate the size of that residual risk?