Project Abstract

The advent of human-level artificial intelligence (HLAI) would pose a challenge for society. The most cost-effective work on this challenge depends on the time at which we achieve HLAI, on the architecture which produces HLAI, and on whether the first HLAI is likely to be rapidly superseded. For example, direct work on safety issues is preferable if we will achieve HLAI soon, while theoretical work and capability building is important for more distant scenarios.

This project develops a model for the marginal cost-effectiveness of extra resources in AI safety. The model accounts for uncertainty over scenarios and over work aimed at those scenarios, and for diminishing marginal returns for work. A major part of the project is parameter estimation. We will estimate key parameters based on existing work where possible (timeline probability distributions), new work ('near-sightedness', using historical predictions of mitigation strategies for coming challenges), and expert elicitation, and combine these into a joint probability distribution representing our current best understanding of the likelihood of different scenarios. The project will then make recommendations for the AI safety community, and for policy-makers, on prioritising between types of AI safety work.