Project guide: Concrete AI paths to influence

Katja Grace and Luke Muehlhauser

Introduction

This document contains suggestions for how researchers could study one set of questions related to "superintelligence forecasting and strategy." We don't have strong reasons to think these are the best methodological choices. You will probably want to modify some of them.

Summary

Compile a list of concrete mechanisms by which an artificial intelligence which is able to outperform humans in any cognitive economic role might accrue influence over world affairs.

Who should do this project?

- Someone familiar with a broad range of computer science and AI/robotics methods.
- This project is likely to be publishable in a narrow range of venues (see 'where to publish')

How this project is useful for superintelligence strategy

A better understanding of the means by which an artificial intelligence might accrue power should allow better prevention of such scenarios, for instance through surveillance or through strengthening security against specific routes. More specific scenarios will also make it more clear whether such prevention is likely to be feasible.

AI risk is also likely to be taken more seriously if there are descriptions of conceivable scenarios in which it takes place, rather than abstract models and hard to imagine scenarios.

Method

1. Collect ideas
    a. See this request for concrete AI takeover mechanisms at LessWrong
    b. Think about it
    c. Ask people who have thought about it. See list 'people to talk to' below.
    d. Fiction has many examples of artificial intelligence takeover (see also mind uploading in fiction)
2. Investigate the plausibility of ideas, and work out some details.
3. Check you have not somehow come upon any suggestions which may be hazardous to publish on (e.g. destructive, non-obvious, and feasible without a superintelligence).
4. Organize them and write a summary of each in an academic style. See Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards for a similar project in a reasonable style.
5. Discuss any overall lessons that appear to you.

Related past work

Request for concrete AI takeover mechanisms (blog post): A request to LessWrong for ideas which could be included in this project.
Artificial Intelligence as a positive and negative factor in global risk details one scenario (See p26).
Superintelligence: paths, dangers, strategies (to be published in 2014) is expected to contain a chapter on strategic relevance of cognitive superpowers, which is likely to be pertinent to this question.
Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards outlines a typology of future risk scenarios. A similar style may work well for this project.
Global Catastrophic Risks is a similar work.
Nanosystems tries to single out plausible future technological scenarios in some detail
A Note on the Confinement Problem is example of detailed description of a problem that future technology would enable. In it, Butler Lampson identified the confinement problem in computer security a couple of decades before an abuse of that vulnerability was detected in the wild.

Where to publish

Some places that have published similar work before include:

- Minds and machines
- Journal of experimental & theoretical artificial intelligence
- AI & society
- Journal of evolution and technology
- Science, technology and human values

Why would they publish it?

The journals listed above might be motivated to publish on this because they are interested in the social implications of technology, or of artificial intelligence specifically.

People to talk to

Nick Bostrom has written about this question, and may have further thoughts on how to approach it, or know of other existing work if there is any.