Introduction to AI Safety

At least the conceptual part

Meta-points

- AGI safety is pre-paradigmatic
- Keep a critical eye
- Keep track of what you think and what the community thinks
- Ask lots of questions
- I’m not an expert

The current evolution and trends

Make them read

https://medium.com/@richardcngo/visualizing-the-deep-learning-revolution-722098eb9c5

I guess

Why be worried? Sounds cool

- Humans have a very general ability to solve problems and achieve goals across diverse domains.
- AI systems could become much more intelligent than humans.
- If we create highly intelligent AI systems, their decisions will shape the future.
- Highly intelligent AI systems won’t be beneficial by default.

What is intelligence?

- Being good at attaining a variety of goals, in a variety of situations.
- There are narrow intelligence (being good at a specific cognitive task) and general intelligence
- There are different notions of powerful AIs: AGI (Artificial General Intelligence), TAI (Transformative Artificial Intelligence) and ASI (Artificial Super Intelligence)

AI Advantages

- Computational resources

- Processing speed

- Duplicability

- Editability

- Improved rationality

Agents and goals

Agency is a spectrum, but here are a couple of important characteristics of an agent:

- Self-awareness
- World-modelling and planning
- Consequentialism
- Coherency
- Flexibility

Instrumentally convergent goals

These are goals that are useful to attain for most utility functions:

- Survival
- Goal-preservation
- Getting smarter
- Getting more material resources

Alignment

An aligned AI is an AI that tries to do what we want it to do.

Another definition is that an AI is aligned if it follows human values.

It is classically divided into two components:

- Outer alignment
- Inner alignment

Useful as an introduction to the problem, but perhaps bad as an operalization of how to solve the problem.

Sources

- Four background claims (Soares, 2015)
- Intelligence explosion: evidence and import (Muehlhauser and Salamon, 2012)
- AGI safety from first principles (Ngo, 2020)