Chapter 13: Choosing the criteria for choosing

The Need for Indirect Normativity

Which values should we attempt to instill in Superintelligence? Difficult question because (1) we are likely to be wrong about many important ethical issues (given our past track record) and, even if we knew the correct ethical theory, (2) we are likely to make a mistake implementing it.

> ### The principle of epistemic deference
>
> A future superintelligence occupies an epistemically superior vantage point: its beliefs are (probably, on most topics) more likely than ours to be true. We should therefore defer to the superintelligence's opinion whenever feasible.

Indirect Normativity: apply this principle to the value-selection problem.

Coherent Extrapolated Volition

Give seed AI the final goal of carrying out humanity's "coherent extrapolated volition" (CEV): what we would want were we better informed about non-moral facts, logically omniscient, free from bias, etc.

Objections: (1) Even if humanity's CEV is well-defined, it would be impossible to find out what it is.
(2) There are so many different ways of life and moral codes that it is not possible for them to all be blended together into one CEV.

Yudlowsky's 4 Arguments for CEV:

1. Encapsulates moral growth
2. Avoids hijacking the destiny of humankind
3. Avoids creating a motive for modern-day humans to fight over the initial dynamic
4. Keeps humankind ultimately in charge of its own destiny

Morality Models

The "Moral Rightness" (MR) Proposal: build an AI with the goal of doing *what is morally right* (relying on the AI's intelligence to figure out what that means exactly).

Worry: It might conflict with our CEV in a way that would be very harmful to us.

The "Moral Permissibility" (MP) Proposal: build an AI with the goal pursuing humanity's CEV unless doing so would be morally impermissible.

Worry: Does this help? (No, not really.)

Do What I Mean

How much cognitive work could we offload onto the AI?

Instruction 1: *Do whatever we would have the most reason to ask the AI to do.*

Problems: (1) Leaves too little room for our own desires, (2) Uses technical vocab (“most reason”) that could easily be misinterpreted.

Instruction 2: *Take the nicest action.*

This approach can only work if we ensure that the AI is motivated to interpret these instruction charitably. Following this thought leads us back to the CEV approach.

Component List

Goal Content	What objective should the AI pursue?
Decision Theory	How should the AI make decisions under risk/uncertainty?
Epistemology	What should the AI’s prior probability function be?
Ratification	Should the AI’s decisions be subject to human review?