Chapter 12: Acquiring Values

The Value Loading Problem

Specify a rule or formula that allows the agent to decide what to do in any given situation. What goals (perhaps, encoded by a utility-function) should we give to AI, and how should these goals be represented?

Our (human) goals?	It's very hard to say what these are exactly.
Explicit Representation?	Only feasible for simple goals; human values are not simple.

How else can we ensure that a Superintelligence has human-friendly values?

Evolutionary Selection

A class of search algorithms with two alternating steps: (1) Expand the population of solution candidates by generating new candidates according to some relatively simple stochastic rule (random mutation, recombination); (2) Contract the population of solution candidates by pruning candidates that score poorly when tested by an evaluation function.

Problems:
1. The process might result in a solution that satisfies the formally specified search criteria but not our implicit expectations.
2. High risk of *mind crime* (the solution candidates might have moral value) if the process is like actual biological evolution.

Reinforcement Learning

Reinforcement-learning agents can be made to learn to solve a wide class of problems by programming them to seek to maximize a reward signal (specially designated percepts received from the environment).

Problem: Danger of *wireheading* (the agent alters its reward mechanism directly)

Associative Value Accretion

Rather than specifying complex values directly, we specify some mechanism that leads to the acquisition of those values when the AI interacts with a suitable environment (much in the same way that actual humans acquire their values).

Problems:
1. Mimicking the value-accretion in humans is unpromising.
2. An artificial value-accretion mechanism might result in the AI acquiring final values different from our own (but, then again, what's so good about *our* values anyway?)
3. The AI, if powerful enough, might disable the value-accretion mechanism.

Motivational Scaffolding

We load values in two steps. First, we give the seed AI an interim goal system with relatively simple final goals that can be explicitly coded. Then, once the AI has developed more sophisticated representational faculties, we replace the interim scaffold goal system with a new set of final goals.

Problem: There's a risk that the AI might become too powerful while running on the interim goal system, preventing programmers from replacing the old goals with the new ones.

Value Learning

Use the AI's intelligence to *learn* the values we want it to pursue. We provide the AI with a criterion that implicitly picks out a suitable set of values. The AI is designed to act according its best estimates of these implicitly defined values (which are continually refined as it learns more about the world).

Example: "Maximize the realization of the values described in this envelope."

But how should we specify the value criterion?

Problems: 1. Risk of perverse instantiation.
2. Which values should we get the AI to learn (what to write in the envelope)?

Emulation Modulation

Manipulate the motivational state of an emulation to tweak the inherited goals of the system.

Problem: Research on emulations might be unethical, cause significant harm, etc.

Institution Design

Design appropriate institutions for a composite system. (Use checks and balances to incentive the agents to behave in appropriate ways.)

Problem: Works better (but also in some ways worse?) for emulations than for AIs. A lot of unknowns here.