**Ethics: EXTRA**

Recently, the EU High-Level Expert Group on AI (2019: 12) formulated four ethical principles that AI applications should meet: respect for human autonomy, prevention of harm, fairness and explicability.

**Human Compatible**

The first ethical code for AI systems was introduced by the famed science fiction writer Isaac Asimov, who presented his Three Laws of Robotics in *Runaround* (Asimov 1942). These three were later supplemented by a fourth law, called the Zeroth Law of Robotics, in *Robots and Empire* (Asimov 1986). The four laws are as follows:

1. A robot may not injure a human being or, through inaction, allow a human being to be harmed;

2. A robot must obey the orders given it by human beings except where such orders would conflict with the first law;

3. A robot must protect its own existence as long as such protection does not conflict with the first or second law;

4. (A robot may not harm humanity or, by inaction, allow humanity to suffer harm.)

The standard view regarding the four laws is that they are important but insufficient to deal with all the complexities related to moral machines.

After more than two thousand years of self-examination, we have arrived at a characterization of intelligence that can be boiled down to this: Humans are intelligent to the extent that our actions can be expected to achieve our objectives. All those other characteristics of intelligence—perceiving, thinking, learning, inventing, and so on—can be understood through their contributions to our ability to act successfully.

Machines are intelligent to the extent that their actions can be expected to achieve their objectives. Because machines, unlike humans, have no objectives of their own, we give them objectives to achieve. In other words, we build optimizing machines, we feed objectives into them, and off they go.

What if, instead of allowing machines to pursue their objectives, we insist that they pursue our objectives? Such a machine, if it could be designed, would be not just intelligent but also beneficial to humans. So let’s try this: Machines are beneficial to the extent that their actions can be expected to achieve our objectives.

AI research, **before uncertainty** became a primary issue in the 1980s, assumed a **world** that was fully **observable and deterministic**, and **goals made sense** as a way to specify objectives.

Sometimes there is also a cost function to evaluate solutions, so an optimal solution is one that minimizes total cost while reaching the goal. For the car, this might be built in—perhaps the cost of a route is some fixed combination of the time and fuel consumption— The key to achieving such objectives is the ability to “mentally simulate” the effects of possible actions, sometimes called **lookahead search**.

Problem: Lookahead algorithms are incredibly effective for their specific tasks, but they are **not very flexible**.

Solution: 1958: John McCarthy proposed a much more general approach that opens up the black box: writing **general-purpose reasoning programs** that can absorb knowledge on any topic and reason with it to answer any answerable question.

To make knowledge-based systems possible requires answering two questions. First, how can knowledge be stored in a computer? Second, how can a computer reason correctly with that knowledge to draw new conclusions?

Answer: use logic: propositional or 1st order logic.

Problem: Very little of our knowledge is absolutely certain. In particular, we don’t know very much about the future. Ignorance is just an insuperable problem for a purely logical system. Uncertainty means that the “purpose put into the machine” cannot be a precisely delineated

goal, to be achieved at all costs. → **Goals** cannot be rescued by looking for plans that maximize the probability of achieving the goal.

Solution: Instead of a goal, then, we could use a **utility function** to describe the desirability of different outcomes or sequences of states. Often, the utility of a sequence of states is expressed as a sum of rewards for each of the states in the sequence. Given a purpose defined by a utility or reward function, the machine aims to produce behavior that maximizes its expected utility or expected sum of rewards, averaged over the possible outcomes weighted by their probabilities.

The modern idea of rationality as maximization of expected utility is sometimes called **Bayesian rationality**. It assumes that a rational agent has access to a posterior probability distribution over possible current states of the world, as well as over hypotheses about the future, based on all its past experience. Researchers in operations research, control theory, and AI have also developed a variety of algorithms for decision making under uncertainty, some dating back to the 1950s.

The purpose is put into these machines in the form of a **reward function**, and the output is a policy that specifies an action for every possible state the agent could get itself into. Limitation: For complex problems where the number of states is enormous and the reward comes only at the end of the game, lookahead search won’t work.

Solution: AI researchers have developed a method called **reinforcement learning**. RL algorithms learn from direct experience of reward signals in the environment, much as a baby learns to stand up from the positive reward of being upright and the negative reward of falling over. The purpose put into an RL algorithm is the reward function, and the algorithm learns an estimator for the value of states. This estimator can be combined with relatively myopic lookahead search to generate highly competent behavior.

Games such as Go and Dota 2 are a good testing ground for reinforcement learning methods because the reward function comes with the rules of the game. The real world is

less convenient, however, and there have been dozens of cases in which faulty definitions of rewards led to weird and unanticipated behaviors.

The **final category** of agent program I will consider is the simplest: programs that connect perception directly to action, without any intermediate deliberation or reasoning in AI it is a **reflex agent.** Reflex agents, then, implement a designer’s objective, but do not know what

the objective is or why they are acting in a certain way, they cannot really make decisions for themselves. One possible way to create more powerful reflex agents is through a process of learning from examples. **Rules are substituted by examples**. *Then* a supervised learning algorithm processes the examples to **produce a complex rule**.

Let's focus on how such methods fit into the standard model of AI: for deep learning there are actually two ways to understand it, depending on the role that the learned rule is going to play in the overall system. The first role is a purely perceptual role: the network processes the sensory input and provides information to the rest of the system in the form of probability estimates for what it’s perceiving.

The problem comes when we move from a purely perceptual role to a decision-making role.

Chapter 3: AI progresses in future

Chapter 4: misuses of AI

Chapter 5: Gorilla problem, King Mida problem and superintelligence explosion Chapter 6: not-so-great AI debate

Chapter 7:

Principles for Beneficial Machines:

Summarize the approach in the form of three 3 principles. They are intended primarily as a guide to AI researchers and developers in thinking about how to create beneficial AI systems; they are not intended as explicit laws for AI systems to follow:

1. The machine’s only objective is to maximize the realization of human preferences. 2. The machine is initially uncertain about what those preferences are. 3. The ultimate source of information about human preferences is human behavior.

Optimism and Caution→

Chapter 8:

Chapter 9: complications

Chapter 10:

If we succeed in creating provably beneficial AI systems, we would eliminate the risk that we might lose control over superintelligent machines. Humanity could proceed with their development and reap the almost unimaginable benefits that would flow from the ability to wield far greater intelligence in advancing our civilization.