RAISING ETHICAL MACHINES

Applying Reinforcement Learning in the Domain of Ethics

by Marten Kaas
AIs Everywhere

- Various tasks and decisions are increasingly delegated to artificial agents whose actions are intersecting the domain of ethics.

DeepMind algorithm beats people at classic video games

Computer that learns from experience provides a way to investigate human intelligence.

Opinion
Artificial intelligence (AI)

I propose that reinforcement learning (RL) methods possess advantages that make them suitable for application in the domain of ethics.
How might we create machines or AIs that can make ethical decisions?

- Creating artificial agents that acts in ways we would approve of.

Is it possible to approach the creation of ethical machines using existing technology?

- It is possible to create a machine or AI that can act ethically (or not).
- We possess the tools now to begin exploring how this might be possible.
- This investigation is worthwhile!
**Supervised Learning Methods**

- Using a labeled dataset so that an artificial agent can extrapolate or generalize to act in novel scenarios.
SUPERVISED LEARNING METHODS

In supervised machine learning, instead of giving behavioural rules to the system, we provide it with examples of input-output behaviour, hoping that it will be able to generalize from the examples and behave well also in situations not shown in the examples. In our running example, we would give the system many examples of pictures of a floor and the corresponding interpretation (that is, whether the floor is clean or not in that picture). If we give enough examples, which are diverse and inclusive enough of most of the situations, the system, through its machine learning algorithm, will be able to generalize to know also how to interpret pictures of floors never seen before.

- Applying supervised learning methods in the domain of ethics is difficult...
Unsupervised Learning Methods

- Using an unlabeled dataset so that an artificial agent can discover hidden structure and patterns.
**UNSUPERVISED LEARNING METHODS**

- Possess advantages over purely supervised learning methods, but are still poorly suited for application in the domain of ethics.

Figure 1: Fraction of Open Images and ImageNet images from each country. In both data sets, top represented locations include the US and Great Britain. Countries are represented by their two-letter ISO country codes.
RL Methods

- Possess advantages over purely supervised and purely unsupervised learning methods.
  - Uses training information to evaluate actions taken, creating a need for active exploration.
  - Type of goal-directed learning from interaction with an environment.
  - Human generated data is not required.
RL Methods

Temporal-difference learning

Dynamic programming

Value/policy

Planning

Model

Direct RL

Acting

Model learning

Exhaustive search

Monte Carlo
An agent can be trained using RL in virtual communities of moral agents using a reward signal that corresponds to the extent to which the agent is able to harmoniously integrate into these communities.
Training ethical machines is challenging...

- Designing the reward signal.

- Explication Problem.

- Methods based on optimization.
  - Is this how we should be approaching ethics?

- Pure RL is also likely insufficient and can be augmented with un/supervised learning, classical AI techniques or even formalized ethical reasoning.
RAISING ETHICAL MACHINES

RL methods operate in the spirit of Turing’s suggestion that we create an artificial child-mind that can be appropriately educated.