Imperial College London

MSc Machine Learning & Data Science

Applicable Mathematics

Ethics of ML/DS Part I

Week 4: value alignment and control

Christoforos Anagnostopoulos

Value alignment and control

- **King Midas, paperclips and trolleys**
- Measure what matters and manage tradeoff
- Prediction versus optimization and control
- Maintaining human oversight

The cautionary tale of King Midas

- King Midas: “I want everything I touch to turn into gold”. Ends up turning his food and family into gold statues. Careful what you wish for.

- Often told as a story about greed, but, taken literally, it is instead a story about value alignment and being more precise when you request something to be maximized

The paperclip thought experiment

- Nick Bostrom, 2003
- “AI, please make paperclips”. Super-intelligent computer turns Earth into a paperclip factory.
- We forgot to clarify: “without killing all life on Earth”. Common sense, but is that complete?
- How about “without enslaving or exploiting people”, or “without polluting the environment”
- Unanticipated consequences, moral dilemmas.

http://www.decisionproblem.com/paperclips/

“You promised us no more science fiction!”

Unfortunately this is not science fiction. Consider social media:

- Natural for social media to optimize “eyes on screen” time (product usage). This can be proxied by like clicks, “show more” clicks, etc.
- It turns out that if I become outraged at something, I will bother to read all the comments. I might also feel the need to respond[1].
- The more provocative the content, the more time spent on screen. Hence, an AI that maximizes screen time will promote divisive posts. An outrage economy.
- Similarly, computer games will create ”addictive”[2] gaming mechanics..
- Twitter initiative to maximize “healthy conversation”[3]. How do we define “healthy”?

[1] https://www.pnas.org/content/114/28/7313
[2] https://www.webmd.com/mental-health/addiction/video-game-addiction
[3] https://blog.twitter.com/official/en_us/topics/company/2019/health-update.html

Value alignment & control poses three distinct problems

Competing values

The moment we consider more than one loss function, we will start introducing trade-offs that must be managed.

e.g., fairness vs accuracy, or privacy vs accuracy

Unknown values

In complex situations, it can be hard to derive the right set of human values from first principles / universally

e.g., rationing healthcare

Corrigibility

Despite our best efforts, we might still encode our values incorrectly. Retaining efficient human oversight is hence necessary.

e.g., self-driving cars actively passing back control to human driver

Are AI systems “moral agents”?

The distinction between moral agents and ethical algorithms can be understood in a similar fashion to narrow and general artificial intelligence: ethical algorithms know what to do according to preset definitions of ethical values in very well-defined situations, whereas moral agents can perform moral reasoning and act autonomously in complex scenaria.

A good criterion is whether an agent is able to navigate moral dilemmas thoughtfully.

Moral dilemmas are more than just tradeoffs between two competing values: they reveal the fact that what seems like a single value from afar might in fact be very context-specific.

Trolley problems

- Consider a train moving fast down a railway track, with its breaks out of order.
- A junction appears in the horizon, and the train is about to go left on it.
- You hold the key to the junction and can force the train to go right if you want to.
- On the left, a family with three kids is crossing the train tracks. On the right, an old man is crossing the train tracks. Do nothing, the family will die. Turn, the old man dies.

This is known as a trolley (old English for train) problem, and it is one of the most famous thought experiments in moral philosophy. It emphasizes the difficulty of relative value.

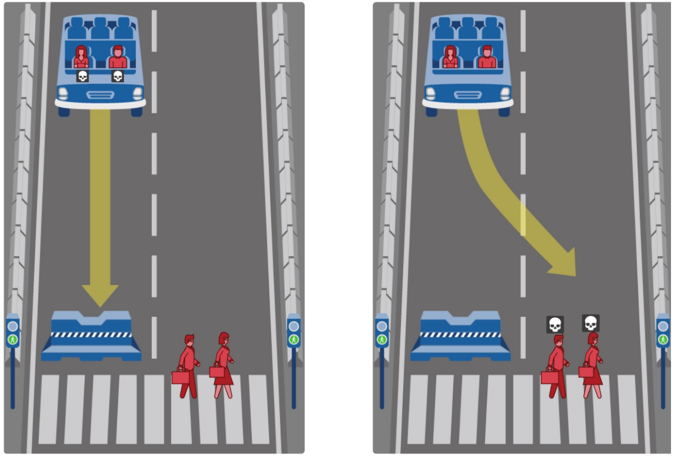
Would you find it harder to actively turn the key than to do nothing in this situation?

Real-life trolley problems

Self-driving cars introduce ***exactly*** this type of dilemma: they have 360 vision, think at the speed of light, and always based on logical reasoning.

What if you have to choose between the safety of the passenger, other cars, and pedestrians?

We will come back to this problem.



https://www.moralmachine.net/

The Car That Knew Too Much. Can a Machine Be Moral? By Jean-Francois Bonnefon

Summary

- The more holistic our approach to measuring risk of harm, the less likely it becomes that we will encounter unanticipated consequences
- In both fiction and folklore we can find cautionary tales about maximizing one thing without properly accounting for side effects and second-degree effects.
- Weighing one harm against another is complex, even for humans.
- Trolley problems are a thought experiment with real-world implications. Generally, choice problems can be used to infer moral preferences.
- A value-aligned, controllable moral agent must be able to deal with competing or imperfectly known objectives, and indefinitely welcome human oversight.