AI Ethics

Dr Janto Dreijer
Ambrite ZA

Zürich - Riga - Stellenbosch

Overview

- What is ethics
- In the news
- Thought experiment
- Two dimensions
  - Data ethics
  - Computational ethics
- Two problems
- Suggestions
- Discussion

Questions to the audience

AI ethics

- Wikipedia: *Ethics* seeks to resolve questions of human morality by defining concepts such as good and evil, right and wrong, virtue and vice, justice and crime.
- Common topics in AI Ethics: existential threat, super-intelligence, value alignment, jobs, discrimination, robots

Some people to follow on AI ethics

- Eliezer Yudkowsky
- Nick Bostrom

AI ethics

- Practitioners have tended to be dismissive of (existential) risks
  - We just want to build cool things
  - Changing
- This is a new challenge
  - Underlying principal components / dimensions
  - Data ethics
  - Computational ethics
- Awaken a sense of urgency
- We’ll focus on near to mid term ethical problems
  - unexpected consequences

In the news: Facebook

Facebook and Cambridge Analytica data breach

- data breach of a collection of personally identifiable information of up to 87 million Facebook users
- data was used to influence voter opinion in Donald Trump’s presidential campaign
- data was detailed enough to create a profile which suggested what kind of advertisement would be most effective to persuade a particular person in a particular location for some political event

In the news: Facebook

- Data ethics: breach of users’ trust in Facebook
- But also implication is that gathered information used *by humans* to guide behaviour of other humans in important ways
- Computational ethics

Thought experiment: “Facebook++”

AI given a task:

- Maximize time user spends engaged with platform
- by changing order of friends’ posts shown
- “Life-style” profile of friends are known

Possible hypothetical maybe long term outcome

- System “figures out” that users that stay at home (pregnant / unemployed) use platform more
- Guide users towards this behaviour by prioritizing posts of friends that have children or are less hard working
- Works because humans evaluate themselves against peer group

Thought experiment: “Facebook++”

Scenario not unrealistic because

- Murphy’s law
  - Anything that can go wrong will go wrong” (given enough time)
- Fixing issues might go against company’s KPIs / profit incentive
  - Built systems reflect company culture
- System behaviour once deployed is outside our control. We constrain system based on input and output. Advanced systems will attempt to optimize objective function without our direct control, in any way it can.

Problem 1: Computers

- Objective function maximization
- Value alignment
    - The sheer complexity of human value systems makes it very difficult to make AI's motivations human-friendly. Unless moral philosophy provides us with a flawless ethical theory, an AI's utility function could allow for many potentially harmful scenarios that conform with a given ethical framework but not "common sense". (wikipedia)
- Trained model will be representative of training data

Problem 2: Humans

- Human “software” is not secure
    - Email + bitcoin + persuasive AI (blackmail?)
    - Spam pays
    - Human attention as finite and controllable resource
- Human nature
    - Democratization of AI is not all good… how powerful are these libraries?
    - Script kiddies doing it for the lulz
    - “Columbine school shooting of AI”

Suggestions

- AI as a Profession
  - Responsibility to the public good / the individual
  - **Professional ethics** encompass the personal, and corporate standards of behavior expected by professionals. The word professionalism originally applied to vows of a religious order (wikipedia)
  - Courses offered at universities. ~~Stanford’s Persuasive Tech Lab~~
  - Recommend an ethics module - engineering, medicine
- **Truthfulness** / clarity / transparency / congruence / “authenticity”
  - Visibility into the black box
- Discuss and educate ourselves
  - Preventative ethics: difficult process of acknowledging and talking through potential ethical issues. Expressing ourselves as professionals.
  - **What do we choose to build?**
  - Yudkowsky

Questions / discussion

janto@ambrite.ch

AMBRITE