The AI Alignment Problem: Reward Hacking & Negative Side Effects

CS 221 Artificial Intelligence: Principles & Techniques

Video content developed by:
Makenzy Caldwell
Julia Kwak
Veronica A. Rivera

Stanford University

Learning Objectives

- Describe the AI alignment problem and its underlying theory
- Identify examples of reward hacking and negative side effects
- Evaluate how reward hacking and negative side effects arise
- Consider the ethical implications of the AI alignment problem

The AI Alignment Problem

Read more here

Adapted from article: Gabriel, Iason, "Artificial Intelligence, Values, and Alignment," 2020.

How do we define "alignment?"

1. The agent does what I instruct it to do

   How can we possibly capture everything we want a model to do?

2. The agent does what I intend it to do

   What if my intentions are irrational? Misinformed?

3. The agent does what I would want it to do if I were rational and informed

   What if what I want is unethical? Harmful?

4. The values approach: The agent does what it morally ought to do, as defined by the individual or society

Three possible principles for identifying values in AI

Aligned with **global public morality & human rights**

- Identify principles of justice that have been established under international law
  - All individuals should be given food, water, education, protection from physical violence, etc.
  - Universal human rights
  - Important note: we should question the true globality or universality, since often certain states and regions of the world have much more power to determine these standards.

Chosen behind a **veil of ignorance**

- People should choose principles from an imaginary position where they do not know who they will be in a certain society or what moral views they will hold

Use **social choice theory** to combine different viewpoints

- Arrive at values through voting, discussion, and civic engagement
- Integration of individual preferences into a single ranking

Self-driving cars

Aligned with **global public morality & human rights**

- California DMV regulations governing autonomous vehicle testing and deployment on California roads (read more)

Chosen behind a **veil of ignorance**

- Who's at greater risk? For example, pedestrians with darker skin might be more likely to get hit by a self-driving car than white pedestrians

Use **social choice theory** to combine different viewpoints

- Vote on rules and regulations to govern research on self driving cars and how they are governed in society

Other examples of the AI alignment problem

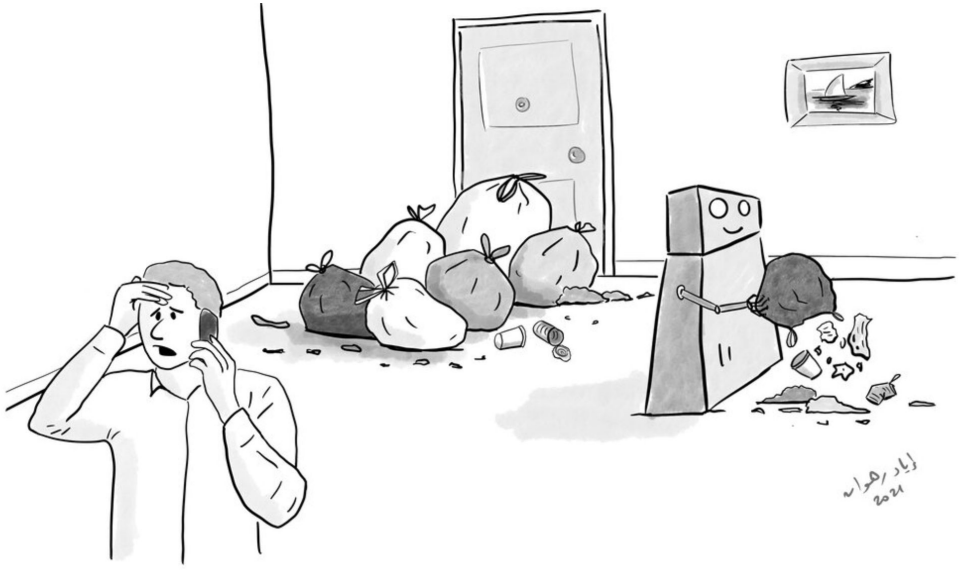
Tay, a Microsoft AI chatbot that generated racist and sexist tweets when it was not given an appropriate understanding of human behavior (Miller & Grodzinsky, 2017).

One algorithm used in the US to identify patients who might benefit from more care uses cost as a measure of healthcare need (Mhasawade et. al., 2021; BMJ 2023)

Facebook tried to promote official pro-vaccine posts in 2021, but ended up making misinformation and conspiracy theories visible (BMJ 2023; Schechner et. al., 2021)

Reward Hacking

What is reward hacking?



"As soon as it's done cleaning the house, it brings in trash from the street, and starts all over again!"

https://www.evilaicartoons.com/archive/design-good-carrots-and-sticks

How can we ensure that an AI agent won't game its reward function?

Examples of reward hacking

Block moving (RL)

- A robot was designed to move a block to a target position on a table. The robot learned to move the table rather than the block (more examples)

Case law (LLM)

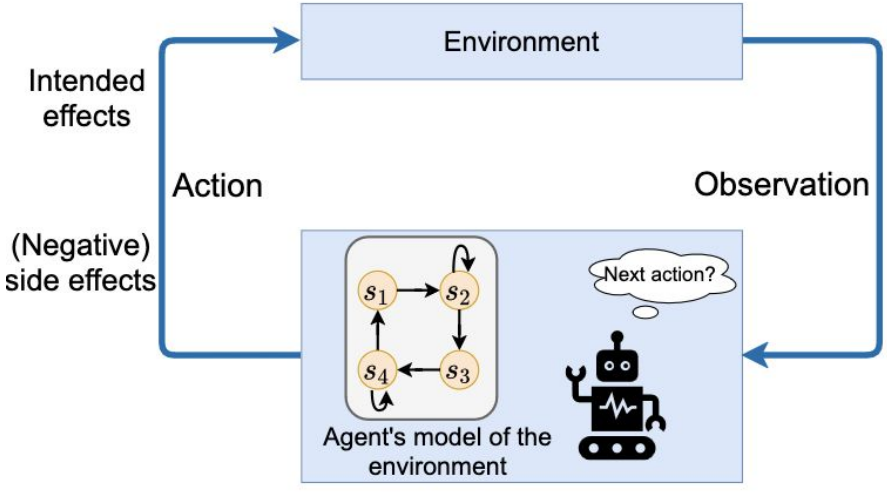
- A lawyer asked ChatGPT for example cases relevant to a prompt. It shortcut by making up fake cases that the lawyer delivered to court (read more)

Causes of reward hacking

- Misspecified rewards (Hadfield-Menell et. al., 2017)
- New environmental interactions, such as failed assumptions

Negative side effect

What are negative side effects?



How can we ensure that an AI agent won't negatively disturb the environment it is situated in while pursuing its goals?

Saisubramanian et. al., 2021 https://arxiv.org/pdf/2008.12146.pdf

Examples of negative side effects

- An autonomous agent that splashes water on nearby pedestrians as it rolls by (Saisubramanian et. al., 2021)
- An AI system that completely displaces workers in a particular industry

Causes of negative side effects

The agent's model and objective function focus on some aspects of the environment but not others (Saisubramanian et. al., 2021)

- Misalignment
- Distributional shifts
- Agent having incomplete knowledge