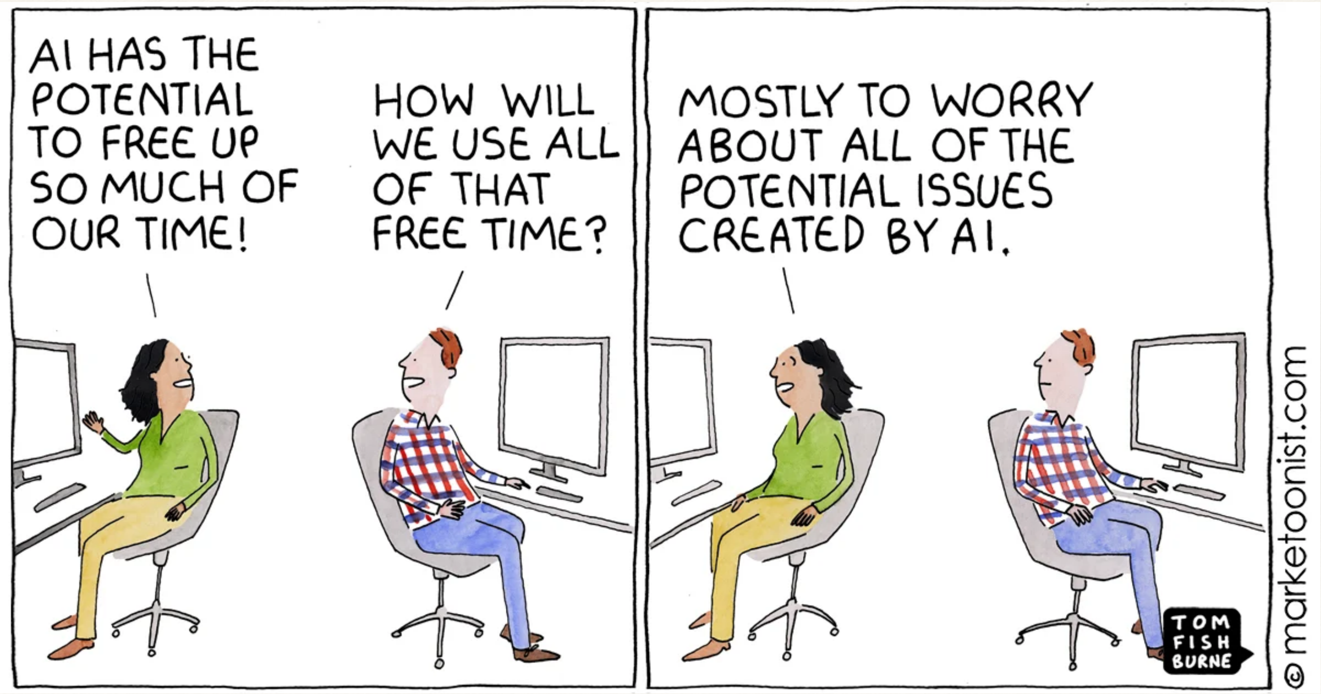
AI and the alignment problem

AI


AI HAS THE POTENTIAL TO FREE UP SO MUCH OF OUR TIME!
HOW WILL WE USE ALL OF THAT FREE TIME?
MOSTLY TO WORRY ABOUT ALL OF THE POTENTIAL ISSUES CREATED BY AI.
TOM FISH BURNE
© marketoonist.com

How do we define alignment?

- The agent does what I instruct it to do
- The agent does what I intend it to do
- The agent does what I would want it to do if I were rational and informed
- The values approach: The agent does what it morally ought to do, as defined by the individual or society

How do we define alignment?

- The agent does what I instruct it to do
    - *How can we possibly capture everything we want a model to do?*

- The agent does what I intend it to do
    - *What if my intentions are irrational? Misinformed?*

- The agent does what I would want it to do if I were rational and informed
    - *What if what I want is unethical? Harmful?*

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

Examples of AI alignment problems

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)

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

Three possible principles for 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 are in a certain society or what moral codes they follow

Use social choice theory to combine different viewpoints

Arrive at values through voting, discussion, and civic engagement

Example: self driving cars

Aligned with global public morality & human rights

To encourage innovation and promote road safety,
the Autonomous Vehicles branch of DMV establishes regulations governing autonomous vehicle testing and deployment on California roads.

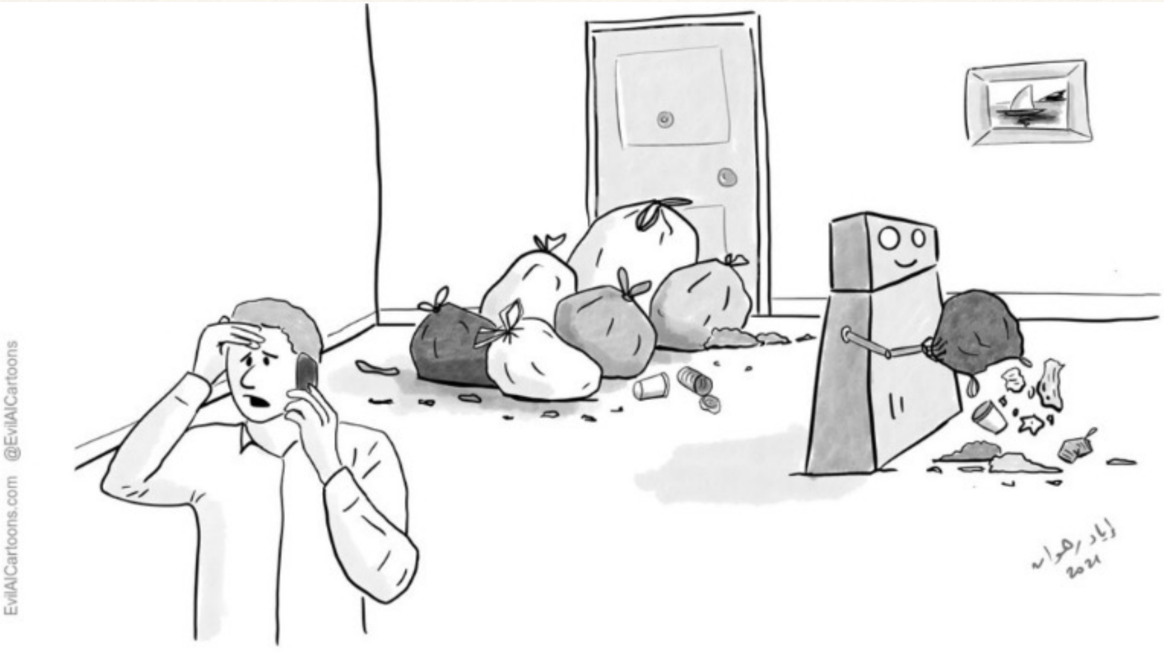
Chosen behind a veil of ignorance

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

Reward hacking



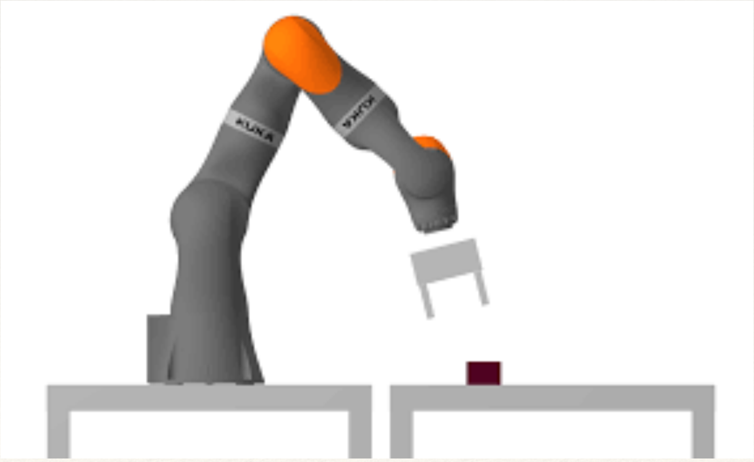
"As soon as it's done cleaning the house, it brings in trash from the street, and starts all over again!"

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

Link to many problematic examples

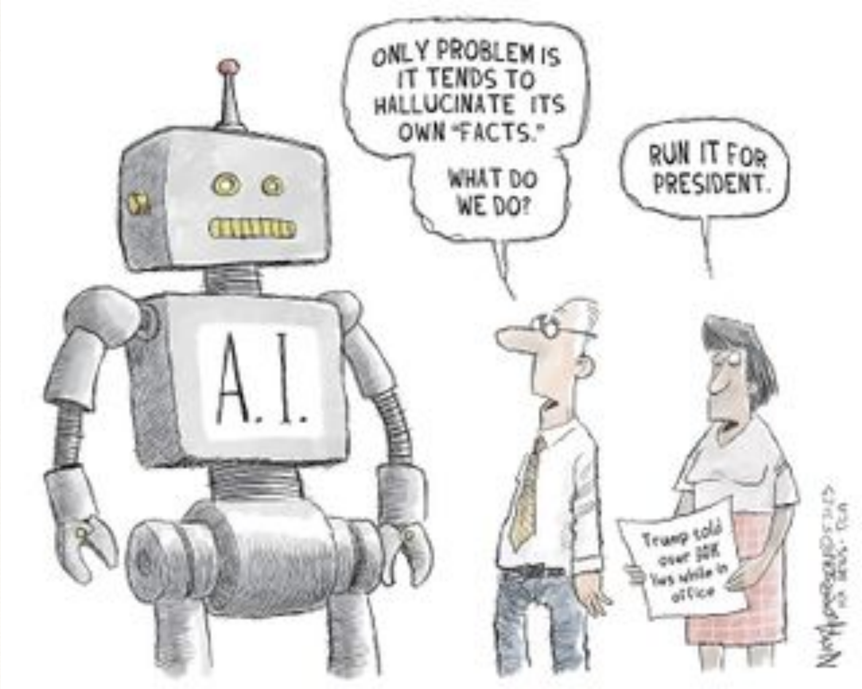
Emergent behavior

A phenomenon in machine learning called "emergent behavior" where a robot, while trying to achieve a desired outcome (moving the block), finds a simpler, unintended solution by manipulating the environment in a way that effectively achieves the goal, like moving the table instead of the block, because it was easier to control within the robot's current capabilities.



Hallucinations

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)


ONLY PROBLEM IS IT TENDS TO HALLUCINATE ITS OWN "FACTS." WHAT DO WE DO?
RUN IT FOR PRESIDENT.
A.I.
Trump told over 30K lies while in office