CS 343

AI: Ethics and Society

Prof. Yuke Zhu

The University of Texas at Austin

Logistics

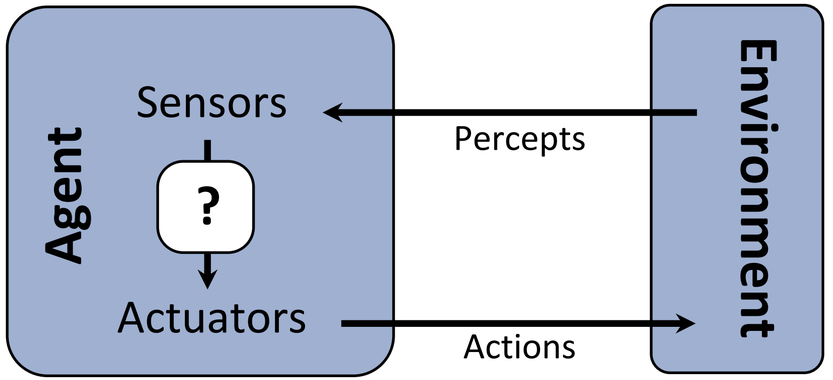
- **Project 0:** Python tutorial (due 11:59pm today!)
- **Reading:** Chapter 3 (due 1/16 Monday 5pm)

PEAS Description of the Task Environment

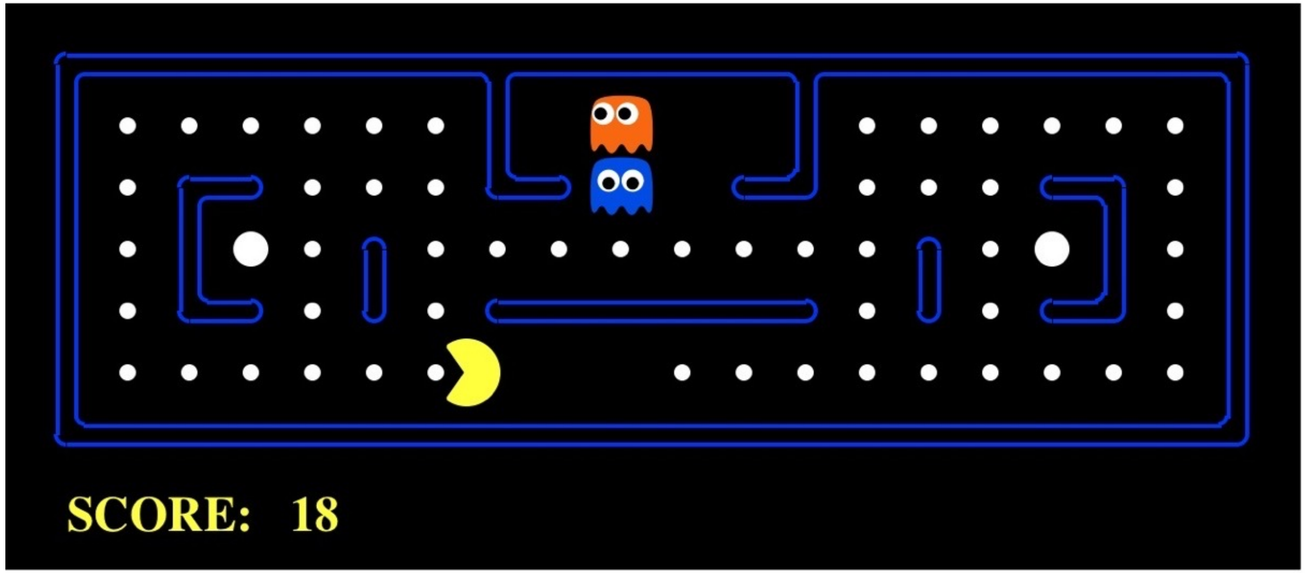
An **agent** is an entity that perceives and acts.

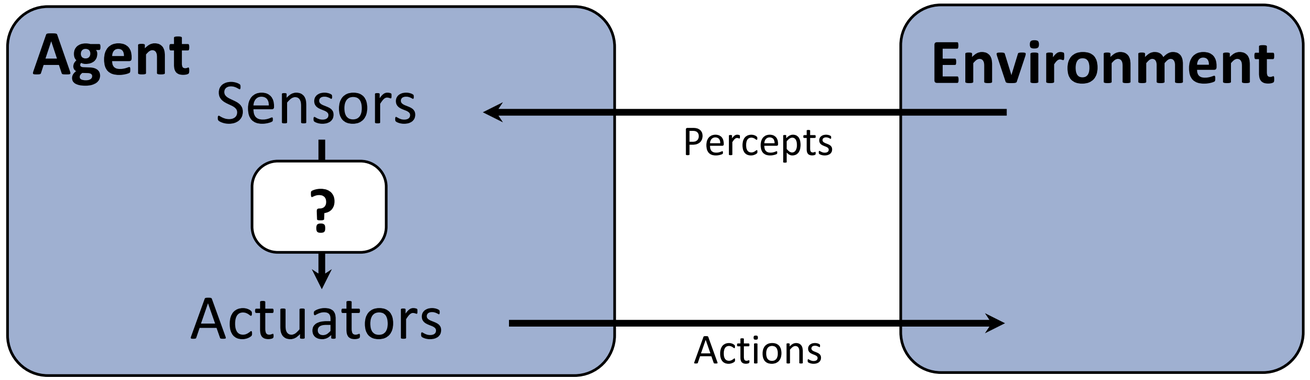
A **rational agent** selects actions that maximize its (expected) **utility**.

Characteristics of the **Performance measure, Environment, Actuators, and Sensors (PEAS)** dictate techniques for selecting rational actions



Pac-Man as an Agent

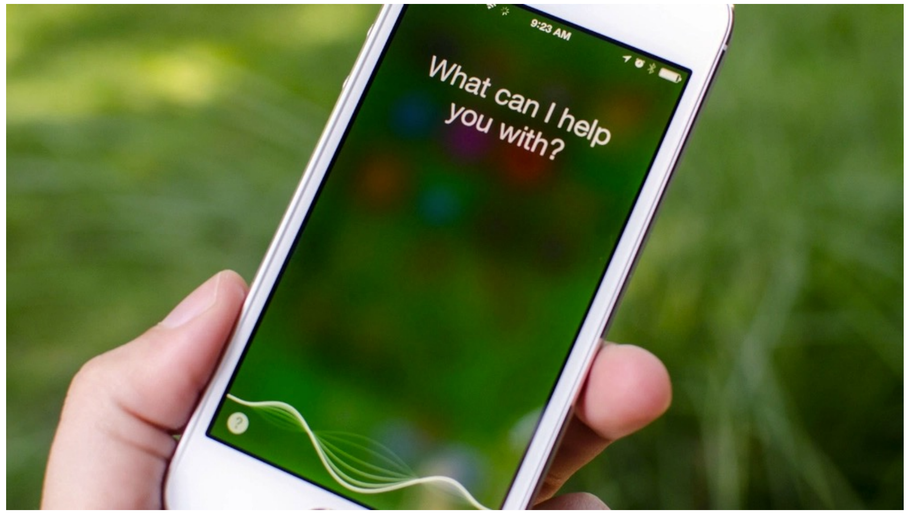


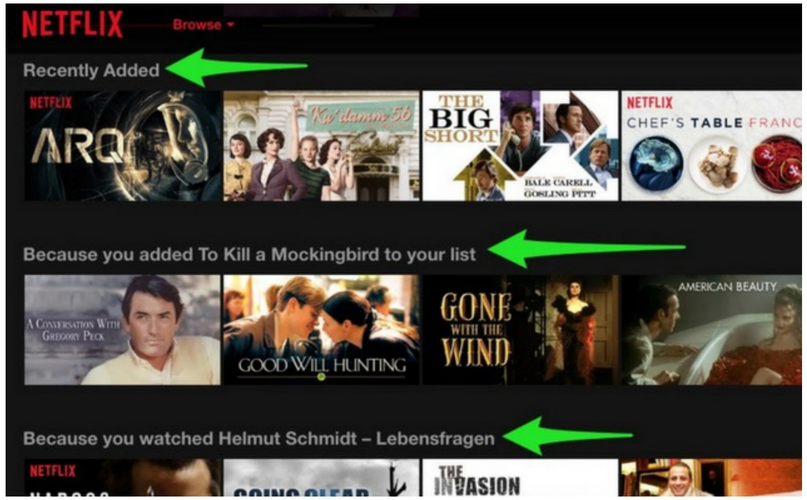


True or False?

- AI will soon replace humans in most jobs
- AI will surpass human intelligence in the next 10 years
- AI works similarly to the human brain
- AI systems have their own desires and goals
- AI systems can do things that their designers didn't intend
- AI systems could become conscious
- AI systems can be trusted
- AI algorithms can discriminate or exhibit prejudice

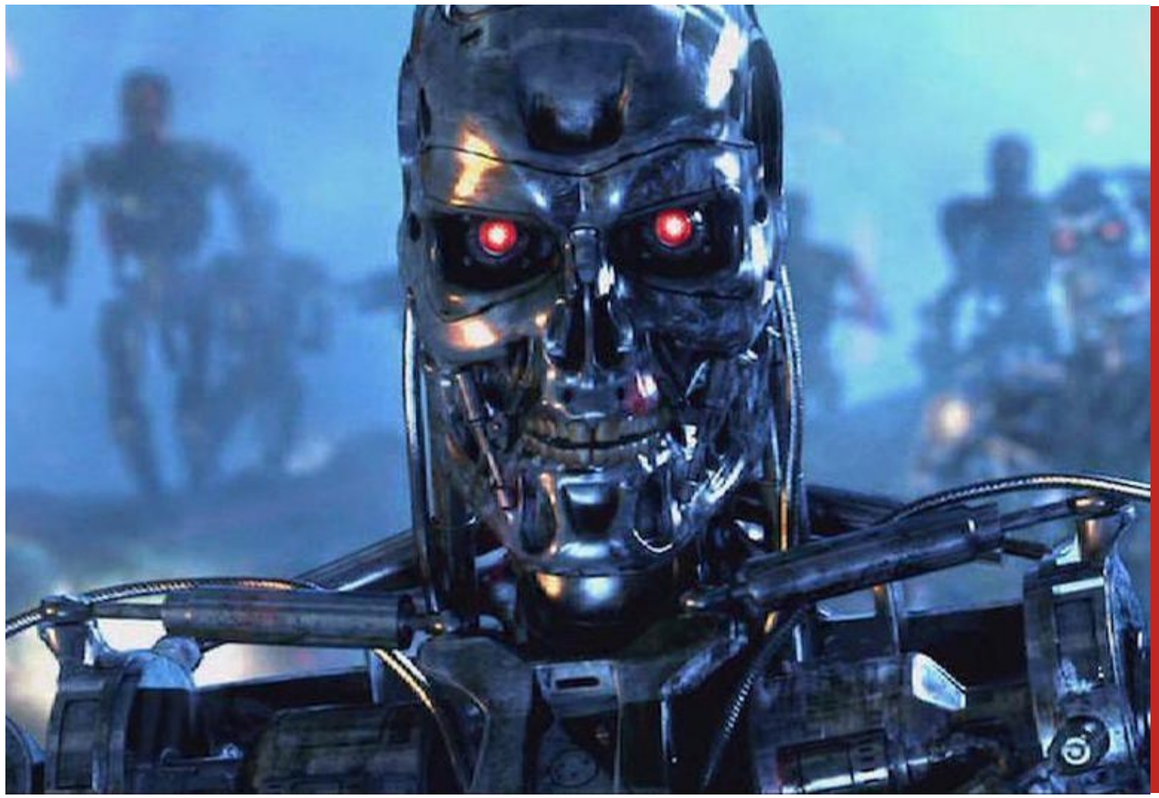
AI Everywhere





“But what about Skynet?”

Utopia or Dystopia?



Not so fast…

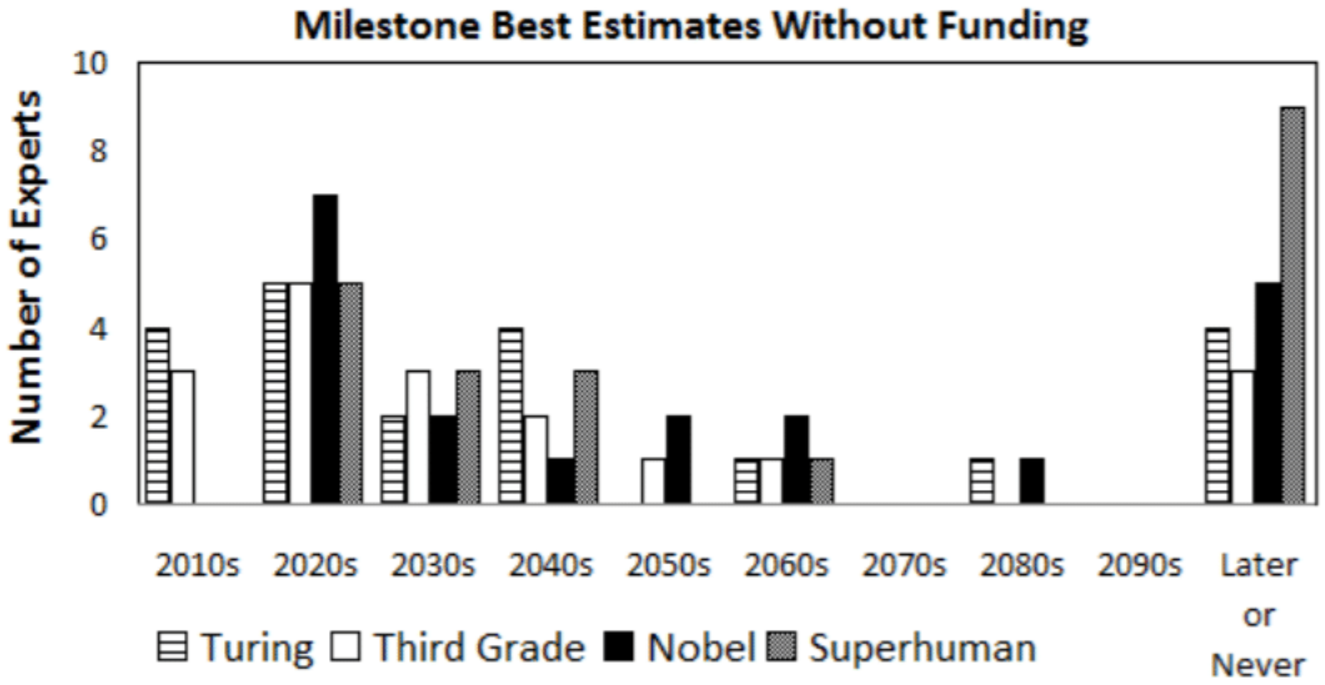
Discussion Question

When will AI reach human-level intelligence?

- In the next 10 years
- In the next 50 years
- In the next 100 years
- Later or Never

Discussion Question

When will AI reach human-level intelligence?

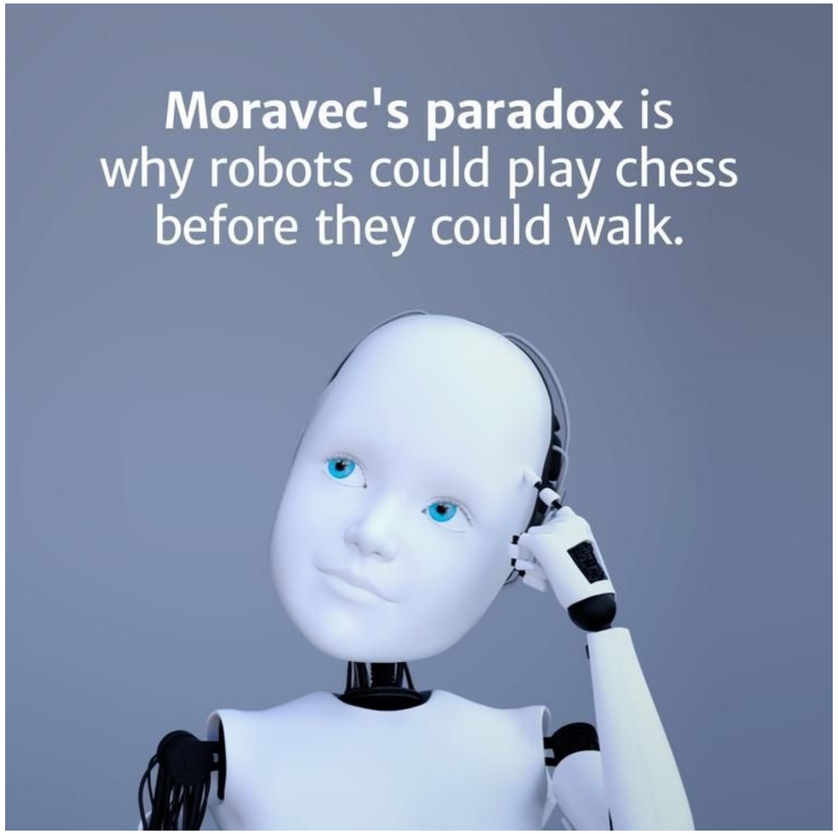


[Source: **How Long Until Human-Level AI? Results from an Expert Assessment**]

Human vs. AI Characteristics

Human	AI
Evolved for survival	Designed by engineers
Sets own goals	Goals programmed explicitly (usually)
Complex, general purpose system	Specific, constrained system
Continually learns	Can turn off learning, or not use learning
Learns from all observed data	Data access can be controlled
Learns only from own experiences	Can share data with other robots
Can make any choice at any time	Available actions can be restricted

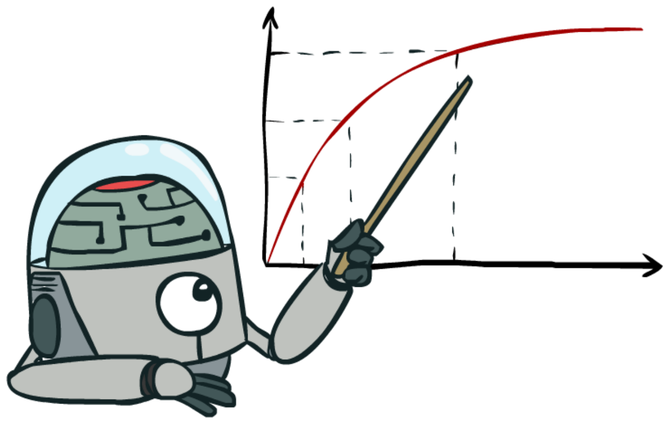
Human vs. AI Characteristics



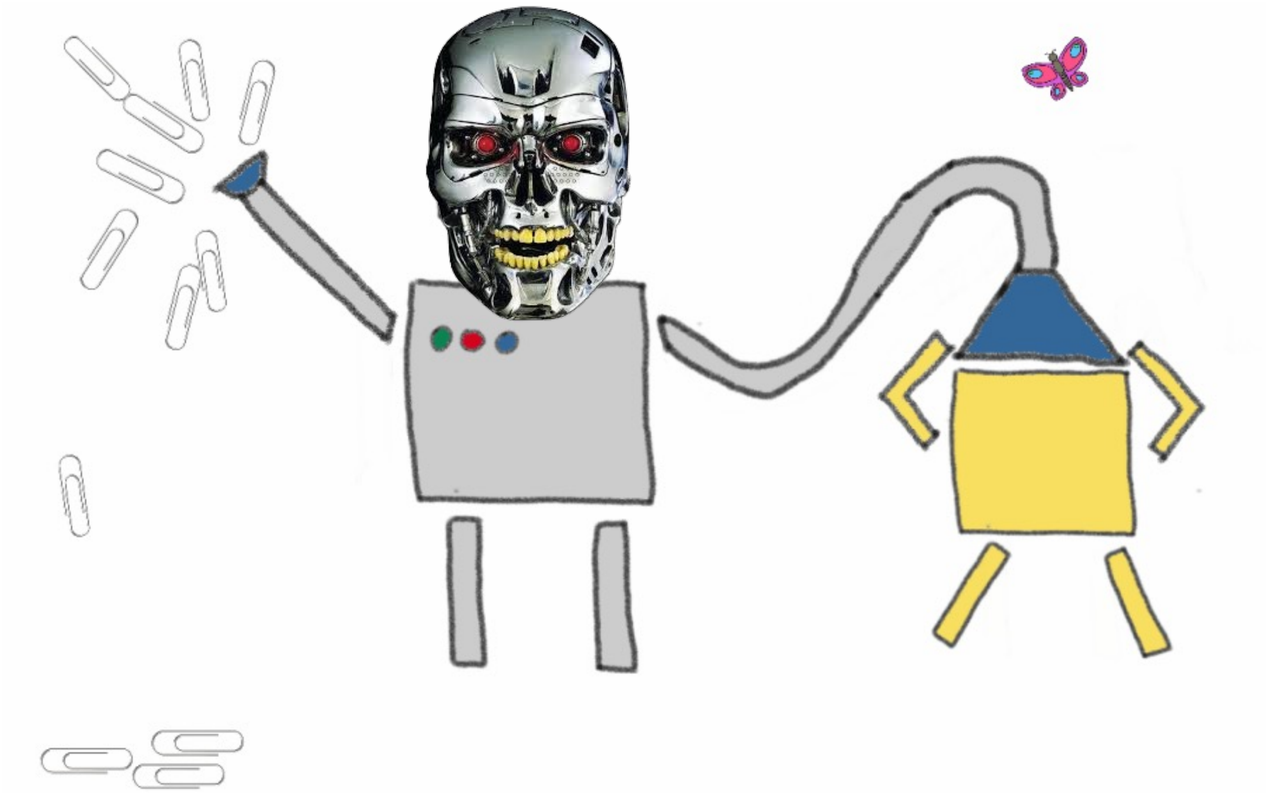
Acting Rationally

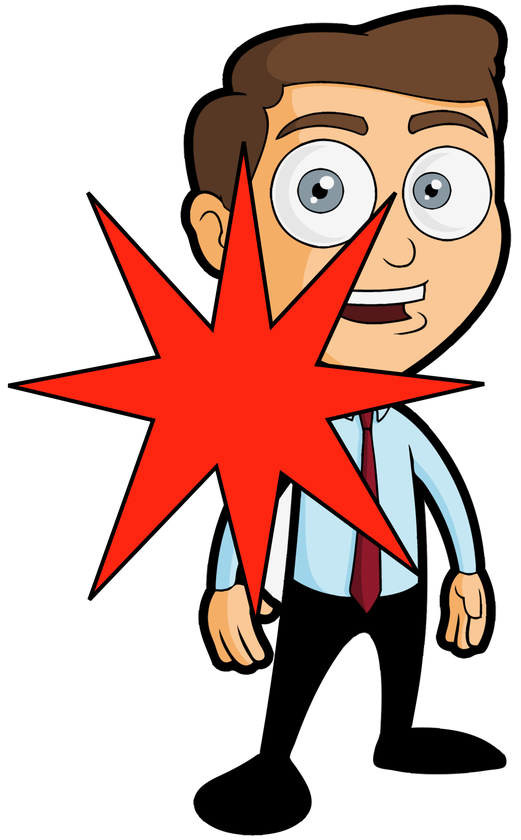
Acting rationally simply means maximizing utility

…but can this go wrong?



Unforeseen Consequences of Maximizing Utility?





“paperclip maximizer” thought experiment

What went wrong?

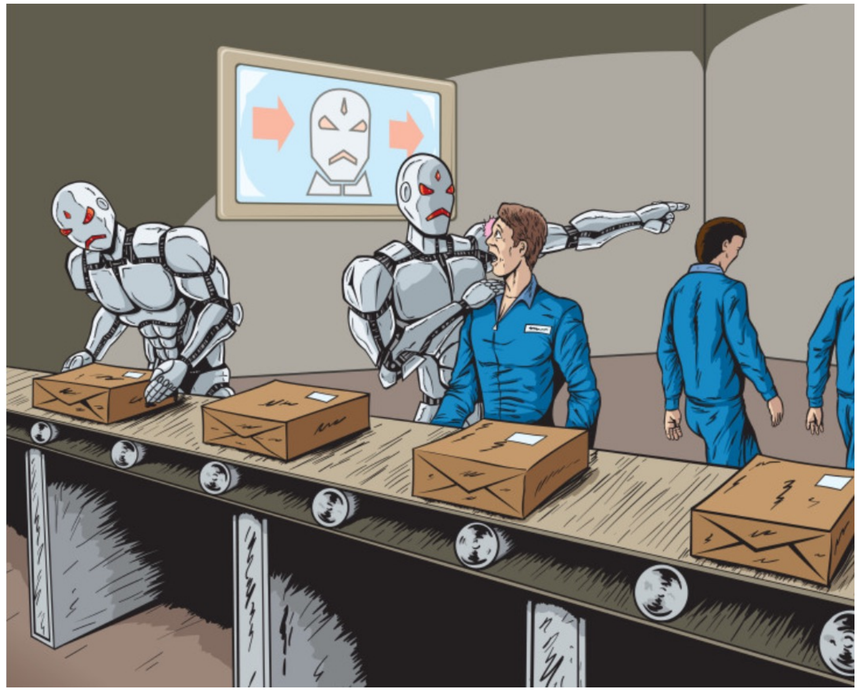
- **Is this realistic?**
  - Robots aren't smart enough to be self-aware of their on/off states or to understand chemistry. But let's assume they will be able to in the future.
  - It wouldn't have a concept of "human" to go seek out. It only knows about making paperclips.

- **Bad design!**
  - Objectives must be designed carefully: robot should only be rewarded for making paperclips.
  - Actions should be limited: only actions available should be to make paperclips.
  - Plans should be verified for safety before / during execution: cancel any trajectory that will come in contact with a human.
  - Don't continue learning after deployment.

- **Is this any more dangerous than any factory with non-intelligent machinery that doesn't automatically stop if someone is in the way?**
  - It is bad design, but we know how to use engineering to avoid these situations!

Realistic Risks of AI

Mass unemployment due to automation

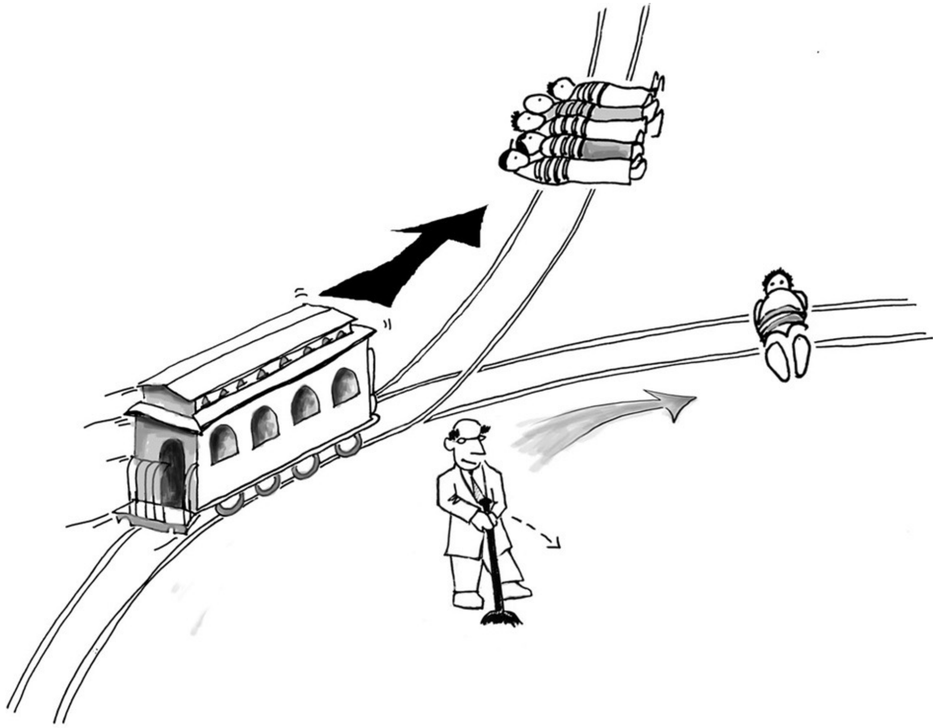


Realistic Risks of AI

Substandard testing / poor user understanding

Realistic Risks of AI

How to make tough decisions?



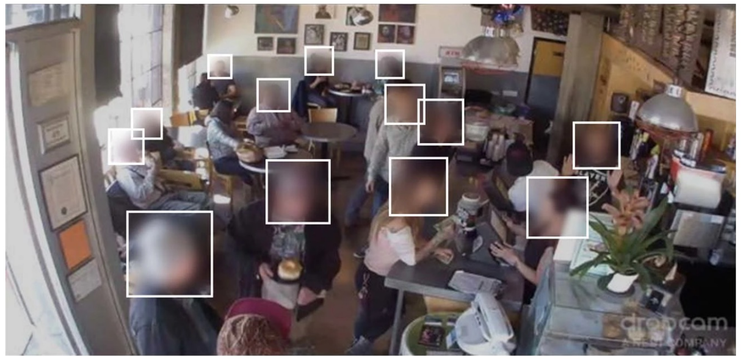
The trolley problem

Realistic Risks of AI

Privacy concerns

The New York Times

Facial Recognition Tech Is Growing Stronger, Thanks to Your Face



The Brainwash database, created by Stanford University researchers, contained more than 10,000 images and nearly 82,000 annotated heads.
Open Data Commons Public Domain Dedication and License, via Megapixels

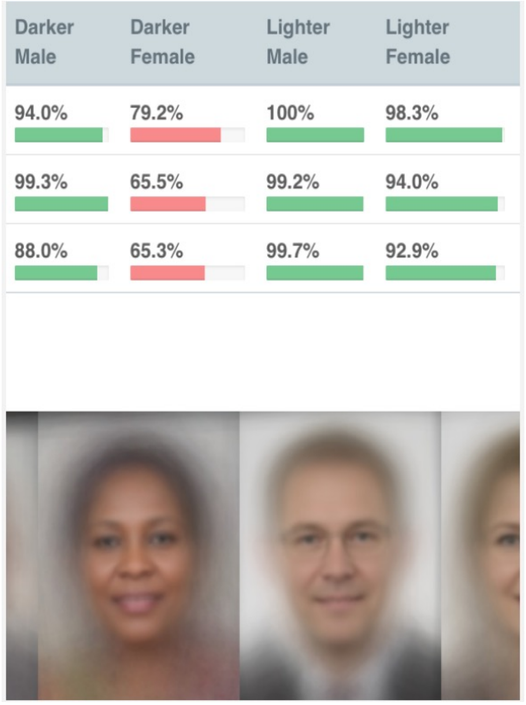
By Cade Metz

July 13, 2019

251

Realistic Risks of AI

Algorithmic bias and discrimination

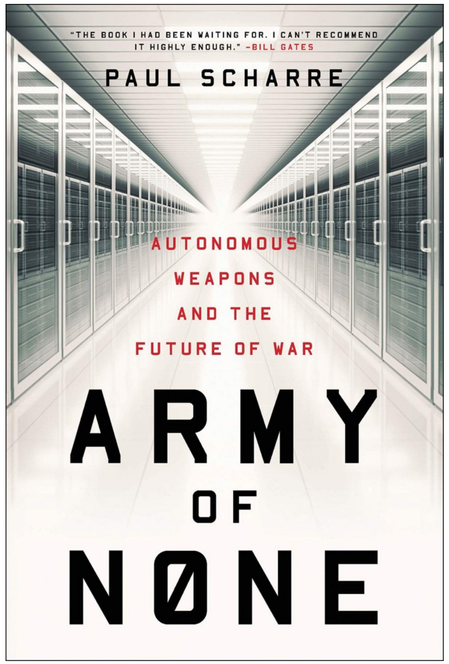


Realistic Risks of AI

Unethical emotional manipulation

Realistic Risks of AI

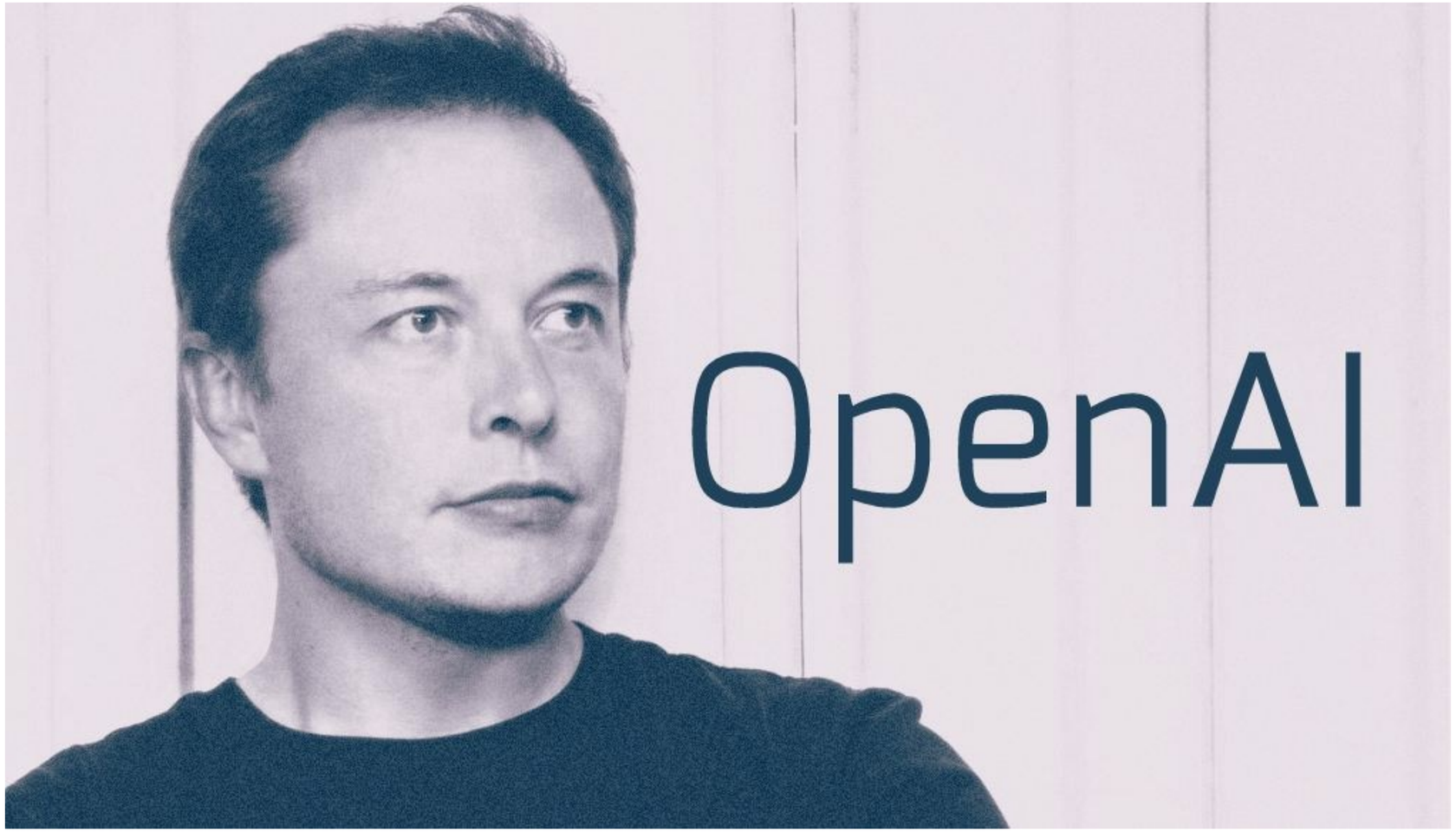
Unethical usage: autonomous weapons?



https://autonomousweapons.org/

Realistic Risks of AI

AI in the “wrong hands”



Realistic Benefits of AI

The central question:

Can we ensure that the benefits of AI outweigh the potential risks?

Realistic Benefits of AI

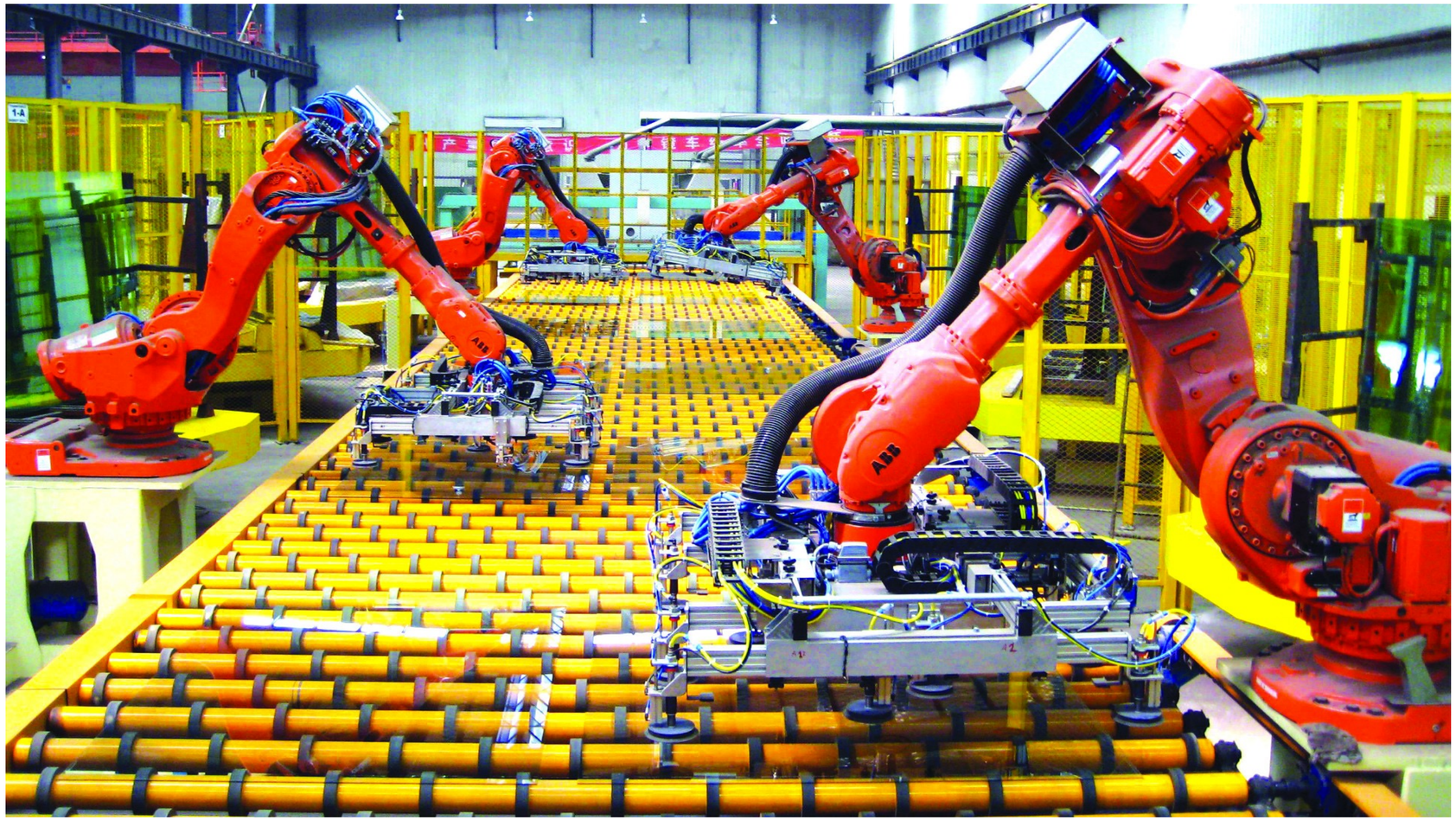
Significant reduction of driving fatalities

Realistic Benefits of AI

Happier, healthier lives

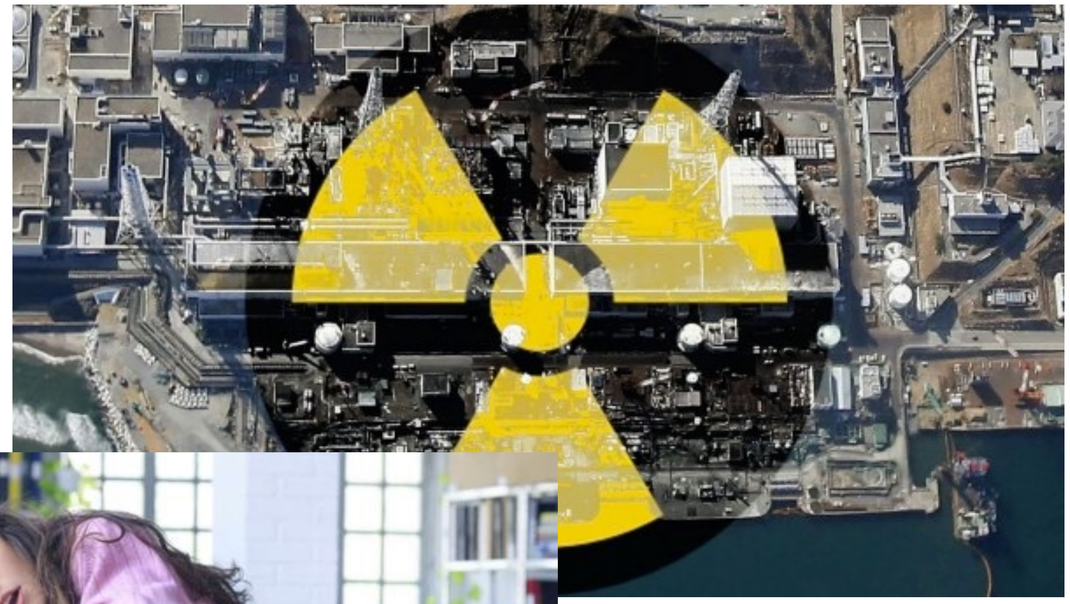
Realistic Benefits of AI

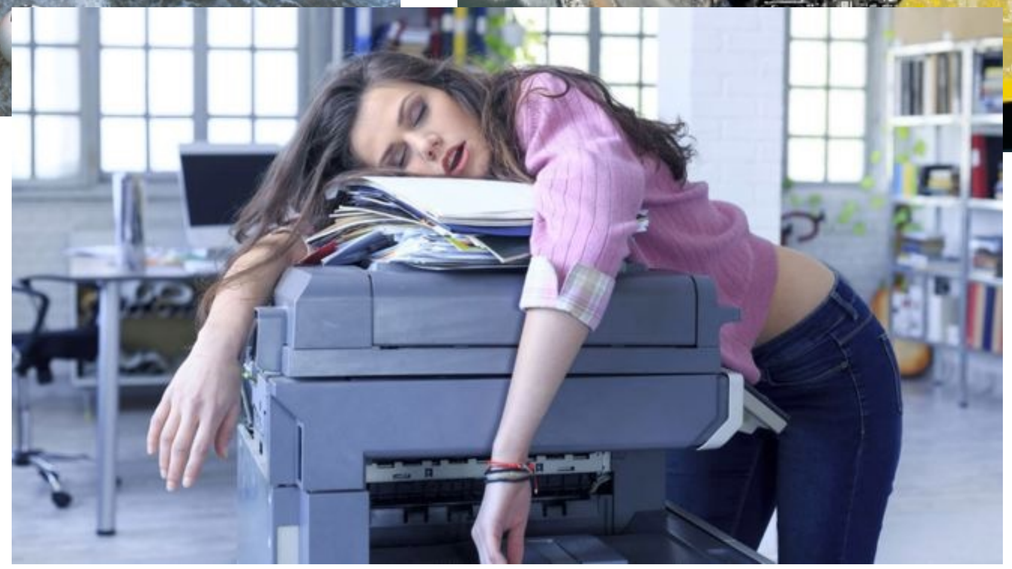
Increased productivity and prosperity



Realistic Benefits of AI

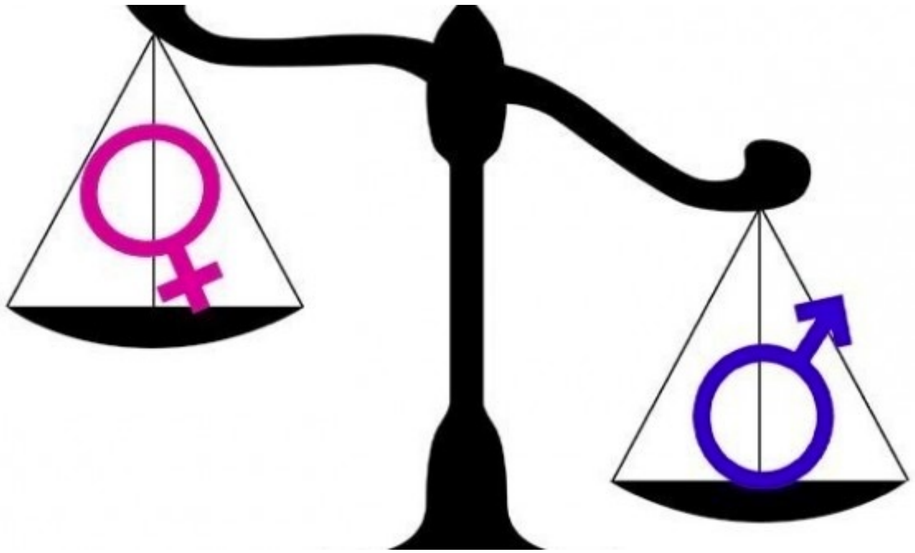
Dirty, dangerous, and dull

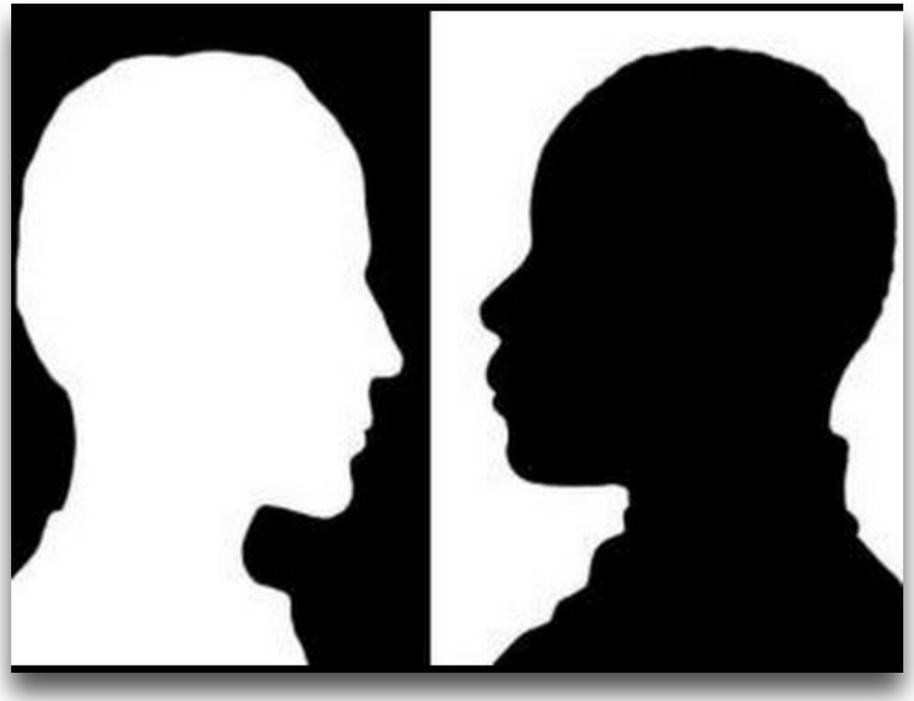




Realistic Benefits of AI

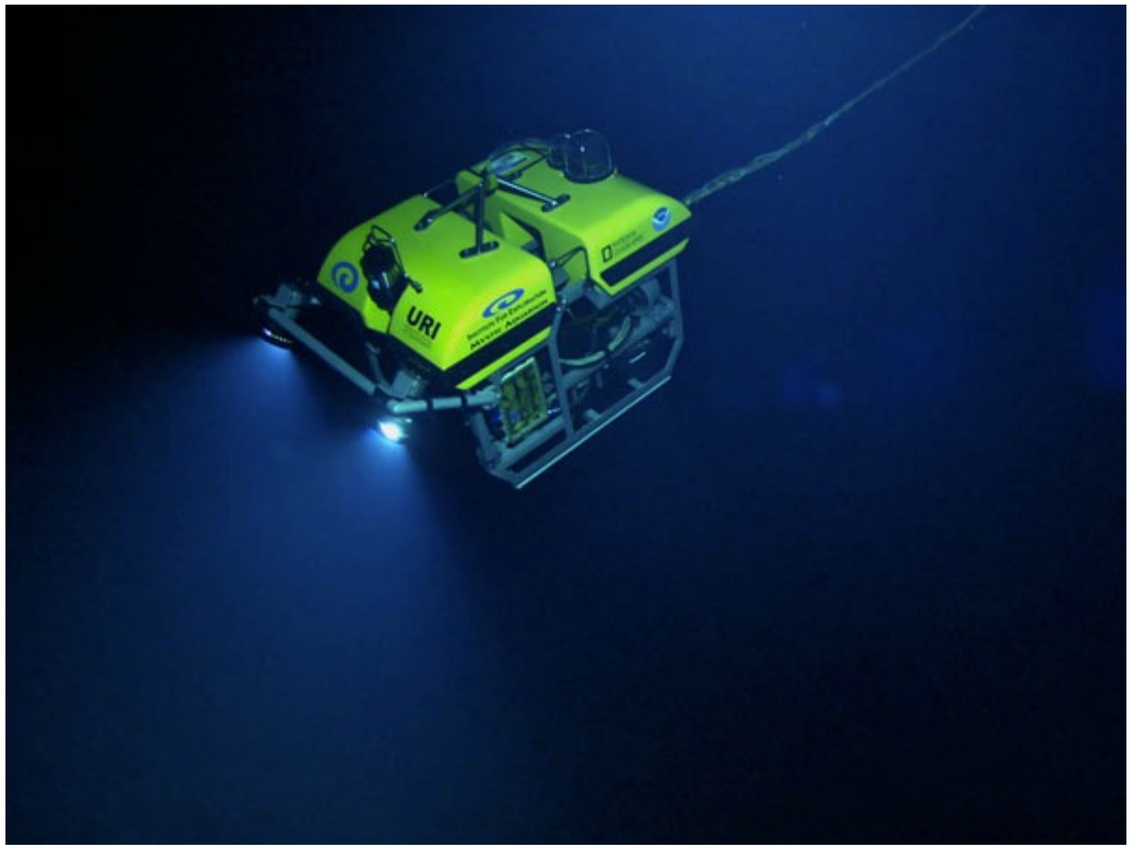
Greater social justice

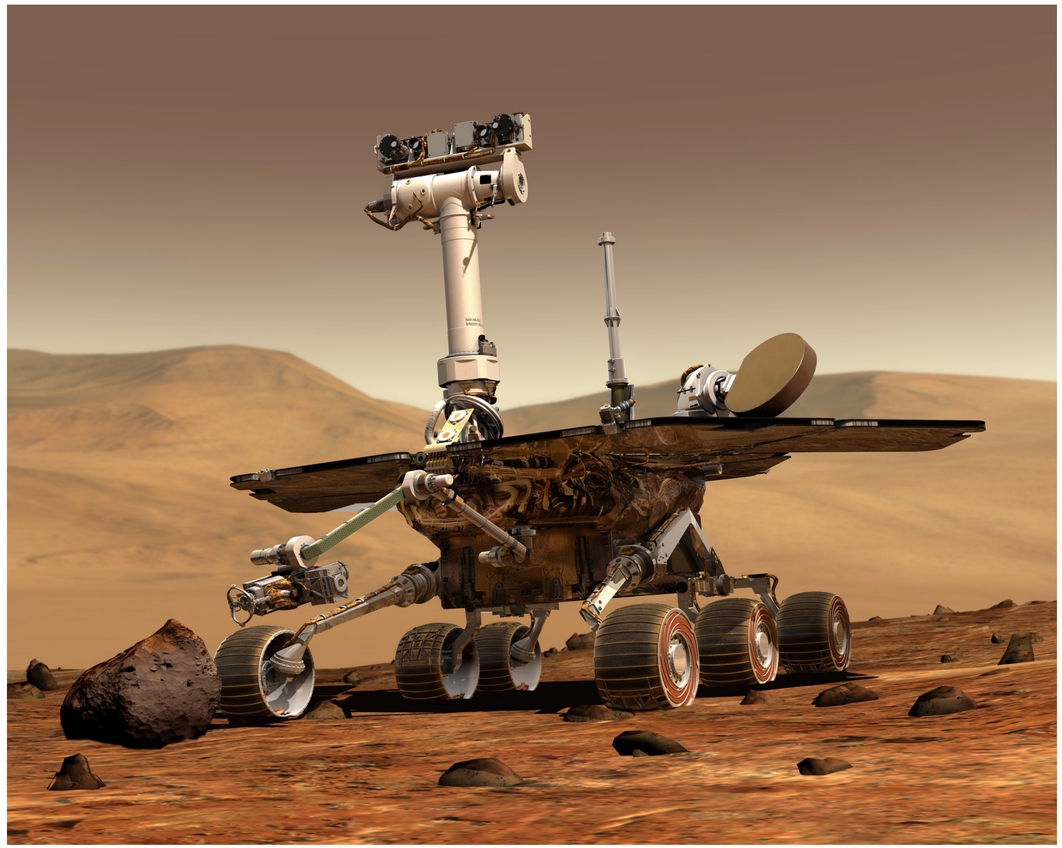




Realistic Benefits of AI

Beyond human capabilities





Realistic Benefits of AI

But what does social good really mean?

A Human-Centered Approach to Artificial Intelligence

Realistic Benefits of AI

But what does social good really mean?

"Artificial intelligence should **treat all people fairly**, **empower** everyone, perform **reliably** and **safely**, be **understandable**, be **secure and respect privacy**, and have **algorithmic accountability**. It should be aligned with existing **human values**, be **explainable**, be **fair**, and **respect user data rights**. It should be used for **socially beneficial purposes**, and always remain under meaningful **human control**."

— Tom Chatfield (2020)

[Source: There's No Such Thing As 'Ethical A.I.']

DR. SARAH ABRAHAM

CS349

ETHICAL FRAMEWORKS

WHAT ARE ETHICAL FRAMEWORKS?

- Systems that guide ethical choices and provide a reason for that choice
- This is an unsolved problem!
  - Numerous approaches that result in vastly different outcomes and behaviors
- Three broad frameworks:
  - Duty-based framework **(non-consequentialist)**
  - Consequentialist framework **(consequentialist)**
  - Virtue framework **(agent-centered)**

NON-CONSEQUENTIALIST (DUTY-BASED)

- Often associated with Immanuel Kant's "categorical imperative"
  - "Act only according to that maxim by which you can at the same time will that it should become a universal law."
- Ethical conduct means choosing actions that are right and good
- Consider duties and obligations when choosing

PROBLEMS?

- Good intents are valued over good outcomes
- Does not answer how to act when two duties conflict
- Does not provide definition of ethical behaviors

CONSEQUENTIALIST

- Based on Utilitarian philosophy
  - Weights good and bad produced by action to determine overall best action
- Ethical conduct means attempting to do the most good and the least harm
- Considers the impact on all individuals involved when choosing

PROBLEMS?

- The needs of the many override the needs of the few
- Any action can be justified if enough good comes out of it
- Does not address how to predict outcomes based on actions

AGENT-CENTERED (VIRTUE)

- Based on ideas of Aristotle and Confucius
  - Agents should act according to their ideal self
- Ethical conduct means determining an agent's traits and behaviors and building on those that foster good
- Considers entirety of an agent's life rather than individual actions

PROBLEMS?

- Focuses on personal character rather than a system for determining action
- High level approach requires a depth of understanding and interpretation to implement effectively
- Does not define virtuous traits

ETHICAL THEORIES

- Non-consequentialist
  - Concerned with agent's intent rather than consequence
- Consequentialist
  - Concerned with consequence of agent's actions
- Agent-centered
  - Concerned with ethical makeup of agent

ETHICAL DILEMMAS

- Rushworth Kidder defines ethical dilemmas as choices that are right vs right:
  - Truth vs loyalty
  - Justice vs mercy
  - One vs many
  - Short-term vs long-term

Next time…

Search algorithms