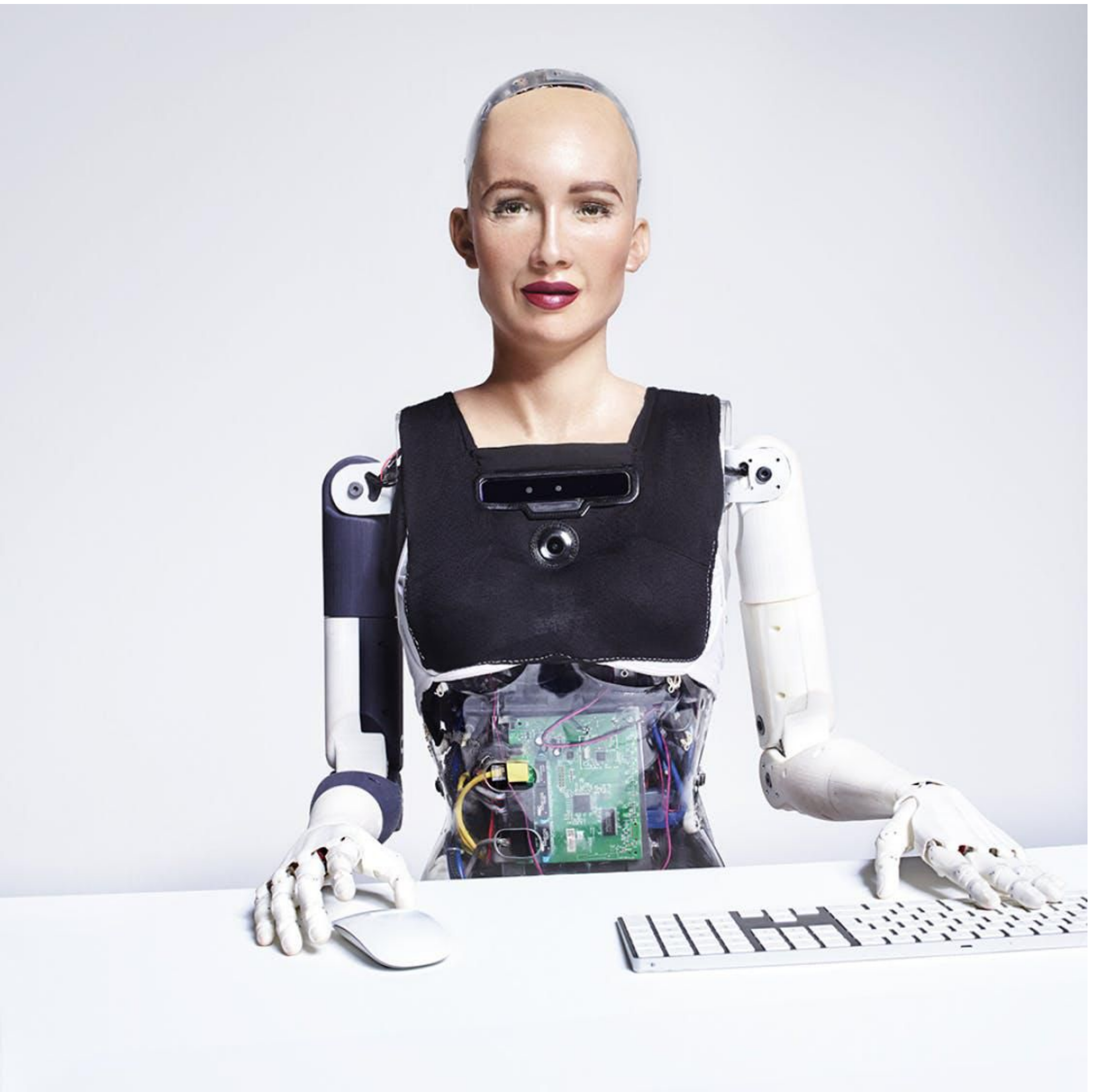
Introduction to Ethics of AI
(main debates)

atoosa.kasirzadeh@anu.edu.au

Sophia: the first Humanoid citizen!



https://www.youtube.com/watch?v=E8Ox6H64yu8

AI systems as objects

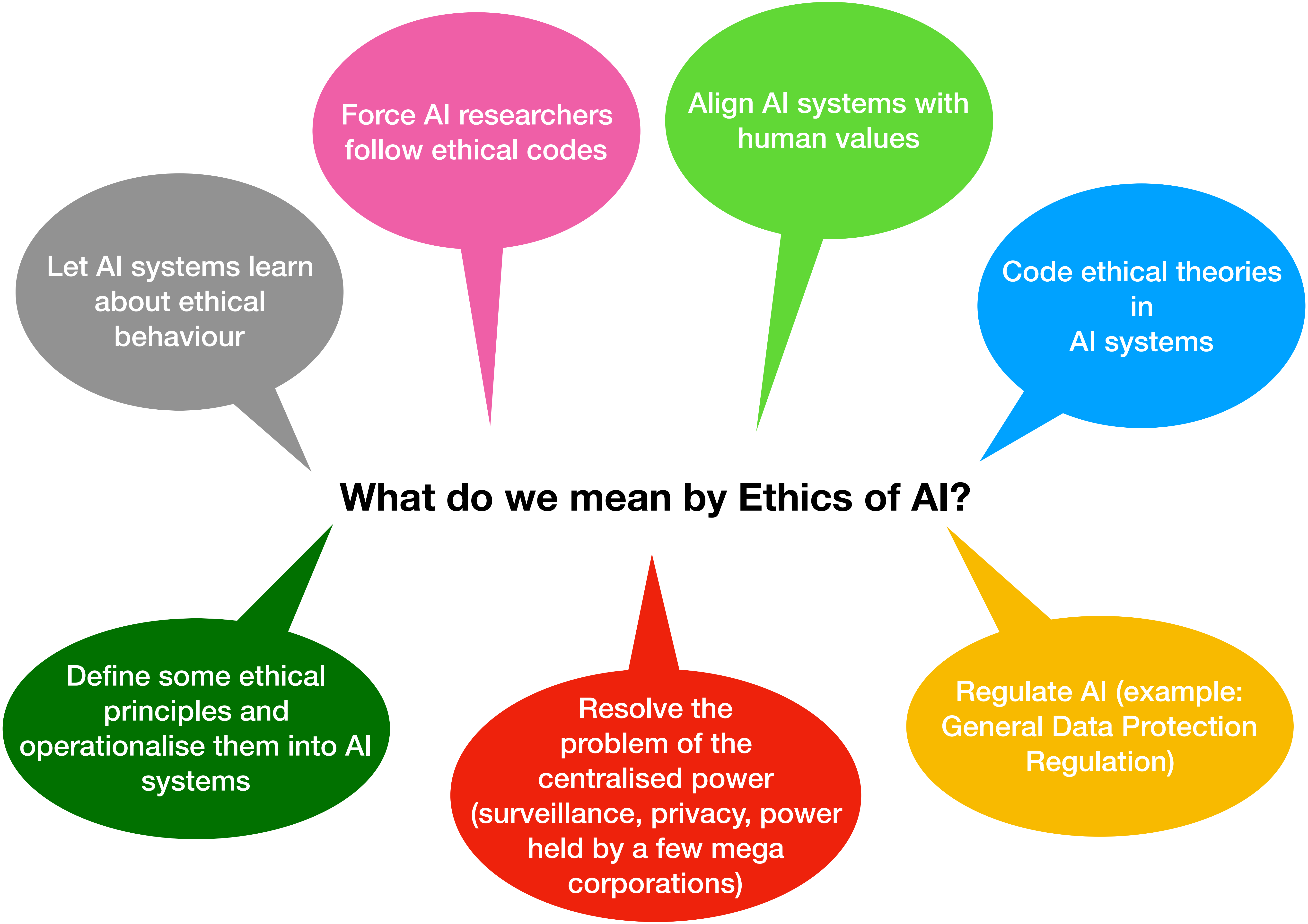
(AI systems are made and used by humans: Ethical issues of the human use of AI)

vs.

AI systems as subjects

(AI systems being the subject of moral agency)

In the case of Sophia, do you think both of these problems are relevant?



What do we mean by Ethics of AI?

Let AI systems learn about ethical behaviour

Force AI researchers follow ethical codes

Align AI systems with human values

Code ethical theories in AI systems

Regulate AI (example: General Data Protection Regulation)

Resolve the problem of the centralised power (surveillance, privacy, power held by a few mega corporations)

Define some ethical principles and operationalise them into AI systems

Code ethical theories
in
AI systems

How to make AI systems Ethical?

Ethical Theories:

what we **ought to do** (what is a right and permissible action to do in each situation)

Motivations —> Actions —> Consequences

- Roughly, different ethical theories belong to 1 of these 3 classes:

1. Consequentialism: Motivations —> **Actions —> Consequences**
   - Through our acts, we shape the world we occupy
   - All that ethically matters concerns what will be brought about

2. Deontological: **Motivations** —> **Actions** —> Consequences

3. Virtue ethics

Utilitarianism (a kind of consequentialism)

- Goodness or badness of alternative courses of action can be measured with some number

- Acting rightly = choosing an alternative with maximal degree of goodness for the greatest number of people

- Classical (hedonistic) utilitarianism: the only thing that is good in itself is pleasure (and absence of pain)

- Preference utilitarianism: the preferences (wants, desires) of sentient beings should be satisfied, to the greatest possible extent

Deontology: **Motivations —> Actions** —> Consequences

- An alternative to any form of consequentialism
- Morality is based on a set of duties or obligations
- Some acts are wrong, even if they lead to the best consequences
- Acting rightly = fulfilling one’s duties
- “**Act** as you would want all other people to **act** towards all other people. **Act** according to the maxim that you would wish all other rational people to follow, as if it were a universal **law**.” (Kant)

Consider the following scenario:

Imagine a murderer comes to your house and rings the bell. Your best friend is at your home. The murderer asks you if you know where your friend is. What is the right thing to do?

Virtue ethics

- Agent-centered rather than act-centered

- Emphasizes the virtues, or moral character, in contrast to the approach that emphasizes duties or rules (deontology) or that emphasizes the consequences of actions (consequentialism).

- A moral person is someone who possesses virtues (as opposed to vices), and show it in their action

- Virtue ethics concentrates on how you can become a better person

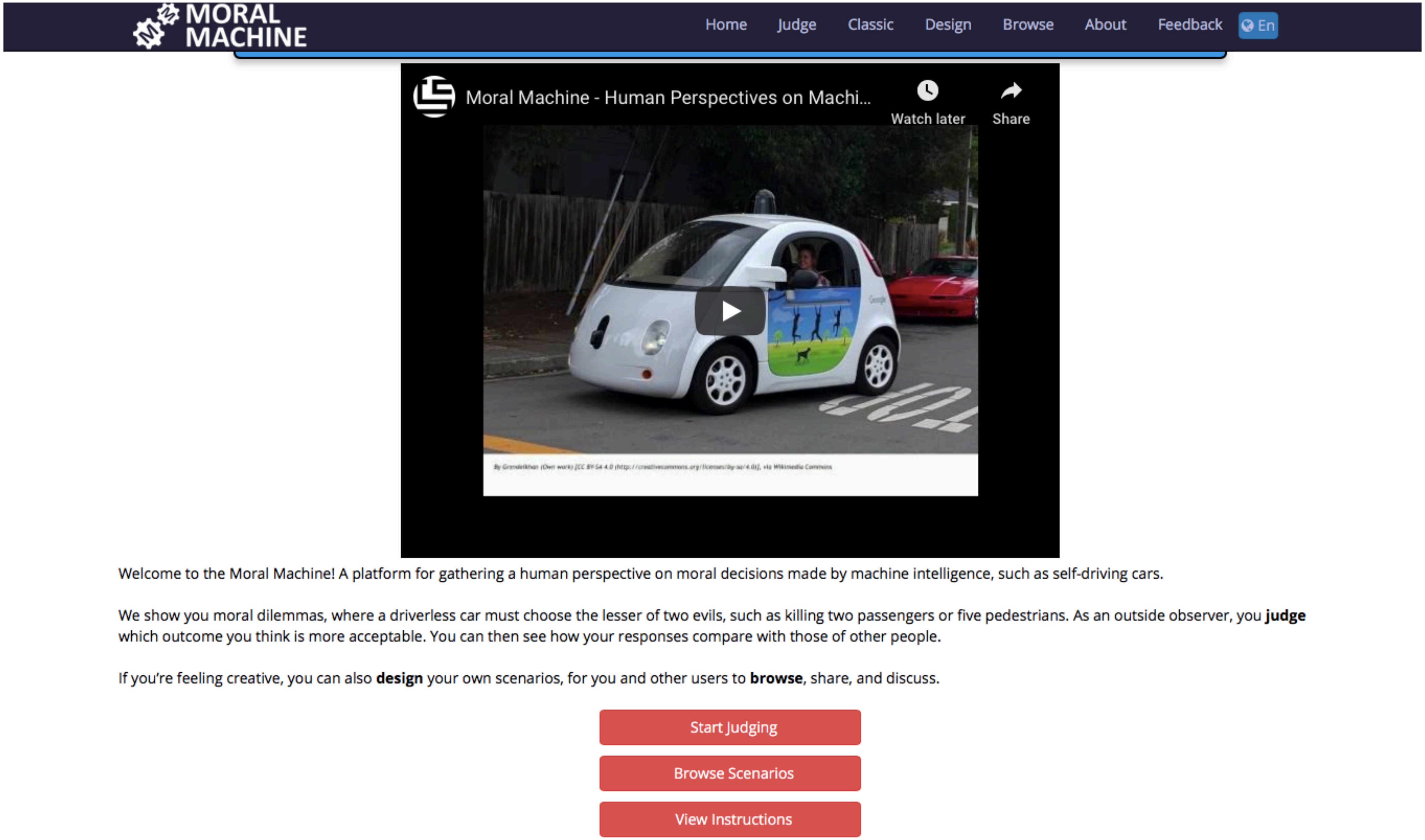
Or Maybe we need some new (non-human) ethical theories:

a new ethical theory (theories) just for machines?

Engineering ethical machines

Top-down strategies : implement (selected) normative theories of ethics and ensure that the moral agent acts aligned with the principles underlying the theory

Bottom-up strategies : ethical theories emerge via the activity of individuals rather than in terms of normative theories of ethics



Welcome to the Moral Machine! A platform for gathering a human perspective on moral decisions made by machine intelligence, such as self-driving cars.

We show you moral dilemmas, where a driverless car must choose the lesser of two evils, such as killing two passengers or five pedestrians. As an outside observer, you **judge** which outcome you think is more acceptable. You can then see how your responses compare with those of other people.

If you're feeling creative, you can also **design** your own scenarios, for you and other users to **browse**, share, and discuss.

MIT Moral Machine Experiment (https://www.moralmachine.net)

(This week's quiz that will be your evaluation of one of these scenarios with respect to the 3 ethical theories that we discussed in class).

Ethical theories & the Ethics of AI

- Three problems:

1. **Theory choice** problem: despite centuries of discussion, ethicists don't seem to agree on what is the right ethical theory
2. **Derivation** problem: it is hard to see how a moral theory provides sufficient information for determining what is right and wrong in practical moral issues, for instance, biomedical ethics (Heyd: "experimenting with embryos: can moral philosophy help?")
3. **Computational complexity** problem

What do we mean by Ethics of AI?

Define some ethical principles and operationalise them into AI systems

Resolve the problem of the centralised power (surveillance, privacy, power held by a few mega corporations)

Privacy & Surveillance

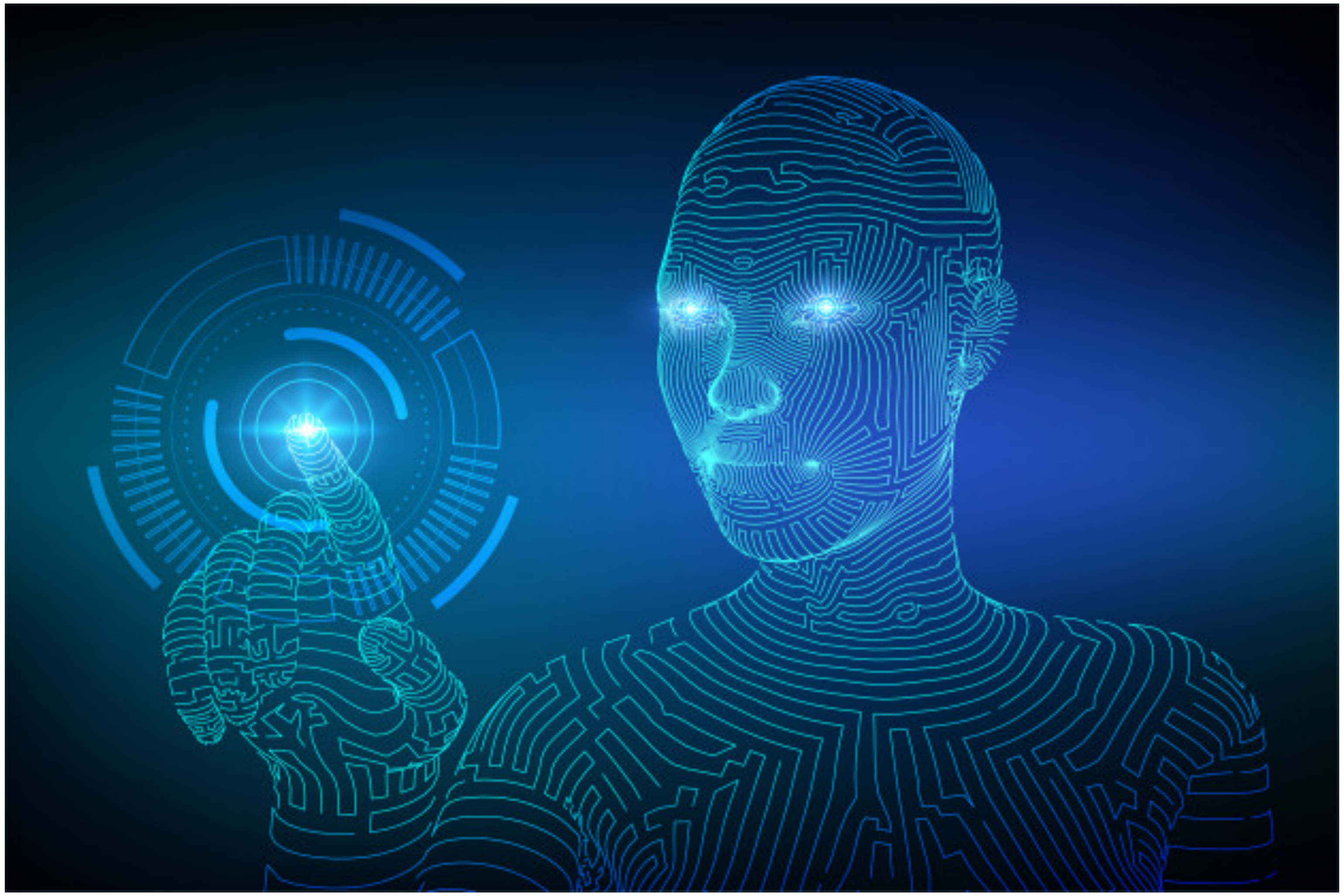
- Concerns about access to private data and data that is personally identifiable
- Privacy as “the right to be let alone”
- Privacy as an aspect of personhood
- Privacy as “the right to secrecy”
- Privacy as “the right to autonomy”
- Surveillance by state or other state agents (businesses, or individuals)

- Data collection and storage are all in the digital sphere

- Most digital data is connected to a single internet

- Sensor technology collects more and more data about all aspects of our lives: knowing more about us than we know about ourselves + having “derived” data

- In the digital sphere, it is harder to control who collects the data and who has access to it: we have lost the control of our autonomy.



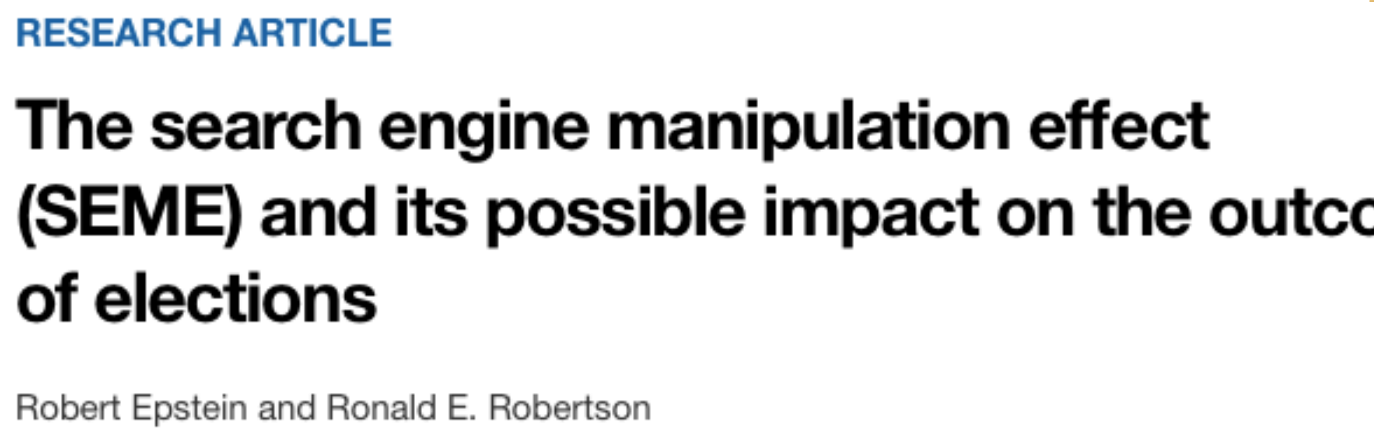
Manipulation of behaviour

- Digital manipulation

- Many advertisers, marketers, and online sellers aim at the maximization of profit

- This maximization can easily require exploitation of behavioral biases, deception and addiction generation

- (?) The search engine (and social media) manipulation effect



RESEARCH ARTICLE

The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections

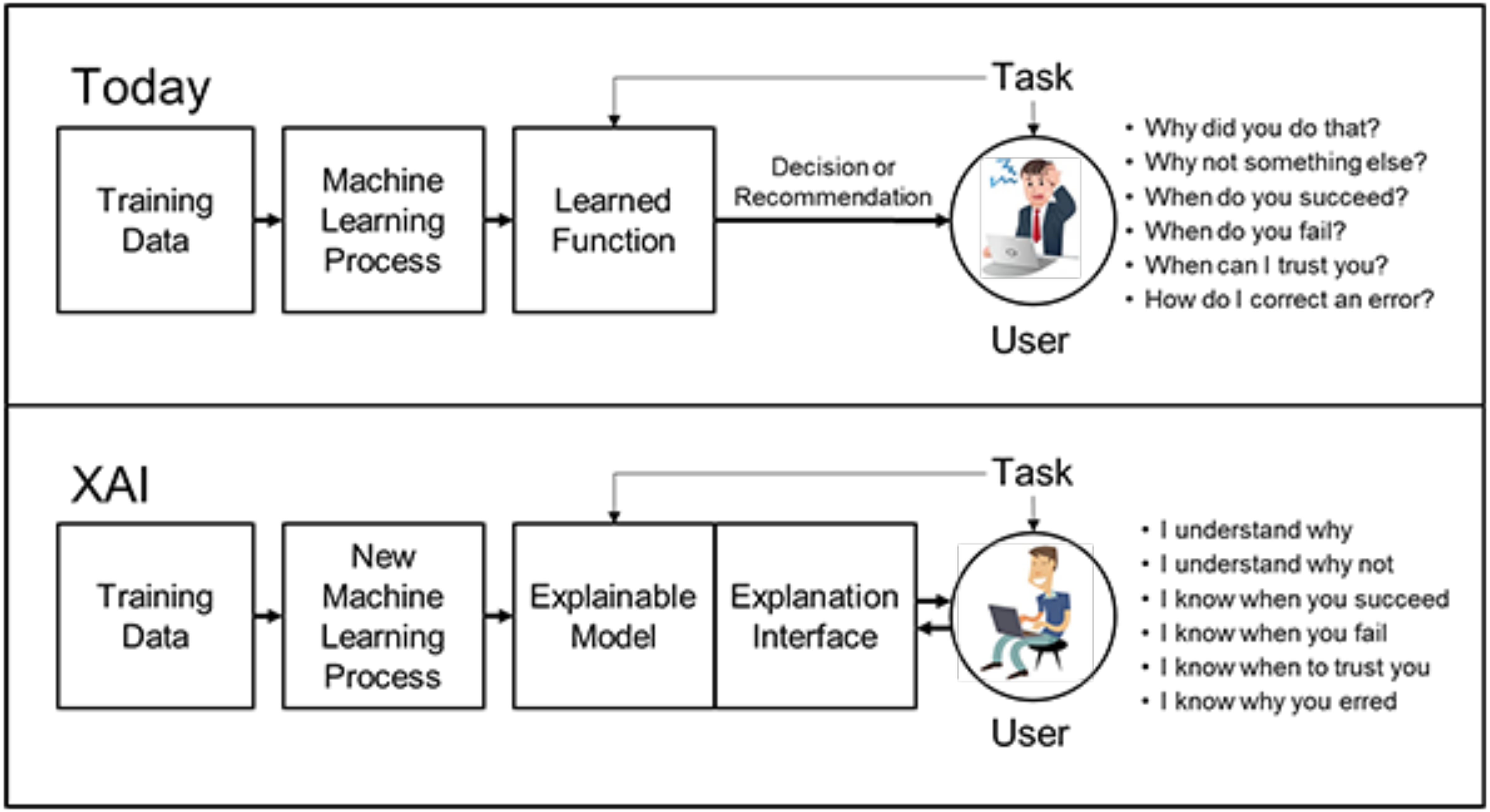
Robert Epstein and Ronald E. Robertson

- Digital manipulation

- Many advertisers, marketers, and online sellers aim at the maximization of profit

- This maximization can easily require exploitation of behavioral biases, deception and addiction generation

Opacity of AI systems



Today

Training Data

Machine Learning Process

Learned Function

Decision or Recommendation

Task

User

- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

XAI

Training Data

New Machine Learning Process

Explainable Model

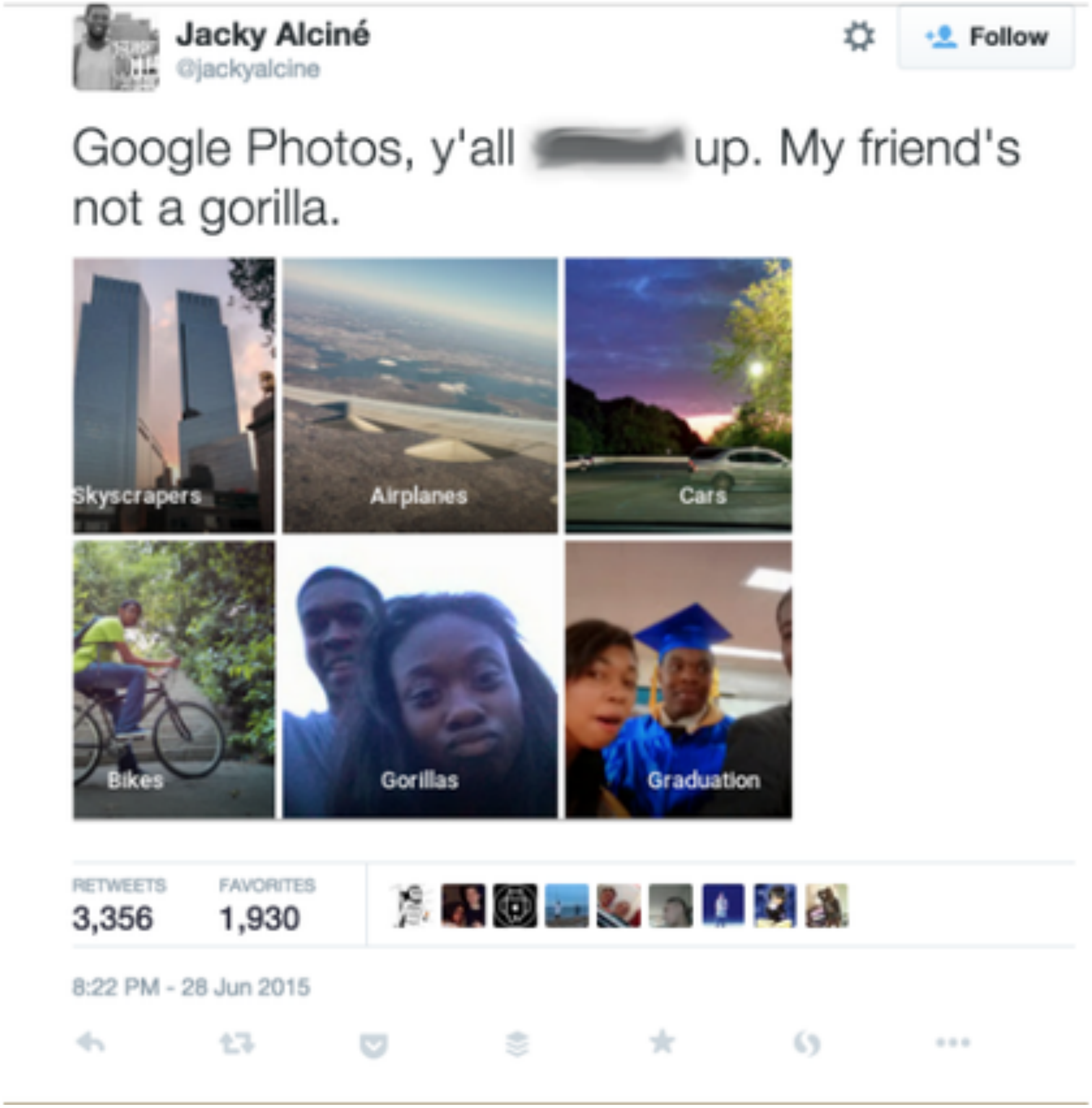
Explanation Interface

Task

User

- I understand why
- I understand why not
- I know when you succeed
- I know when you fail
- I know when to trust you
- I know why you erred

Bias and Fairness in Decision Systems



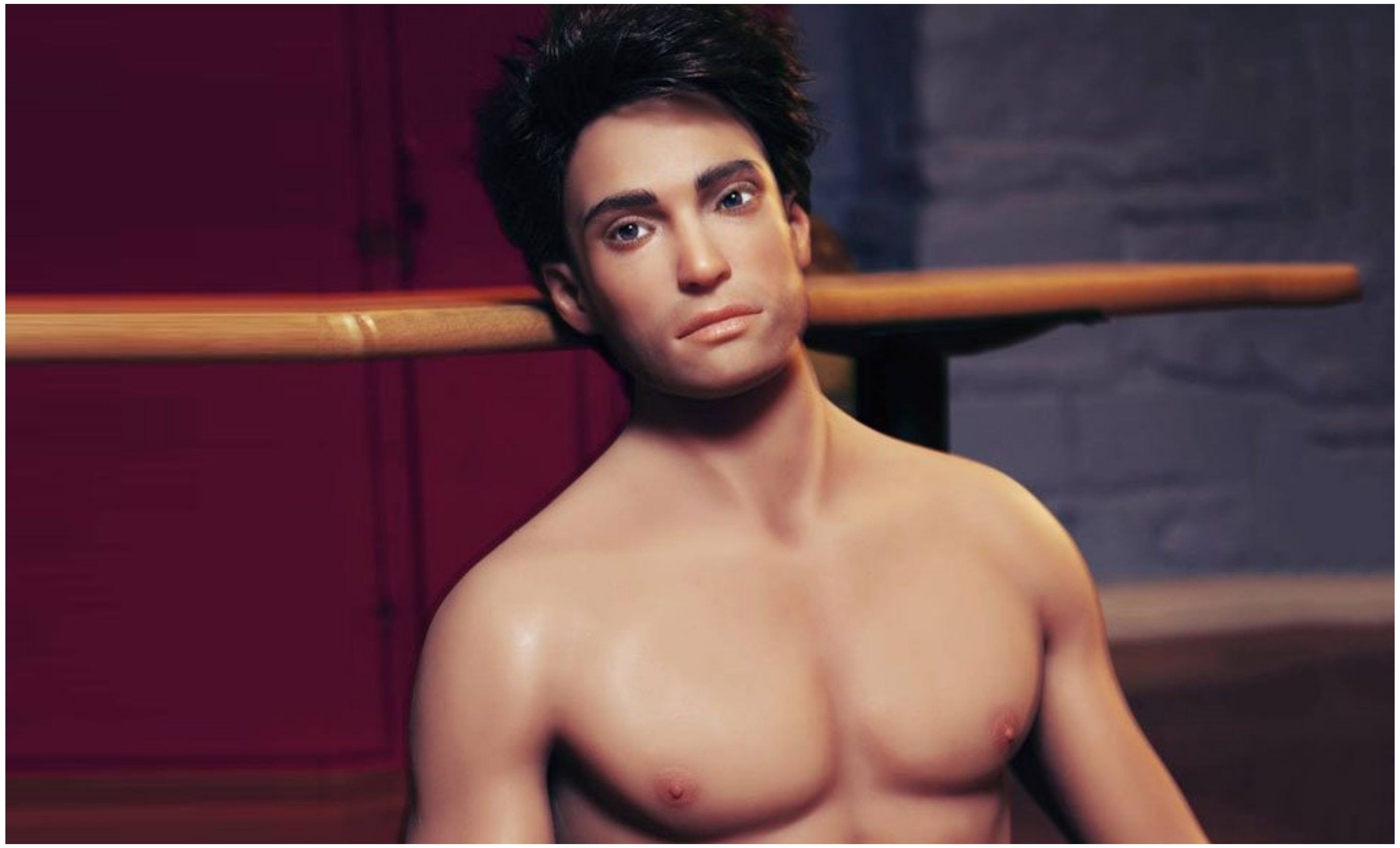
Jacky Alciné
@jackyalcine

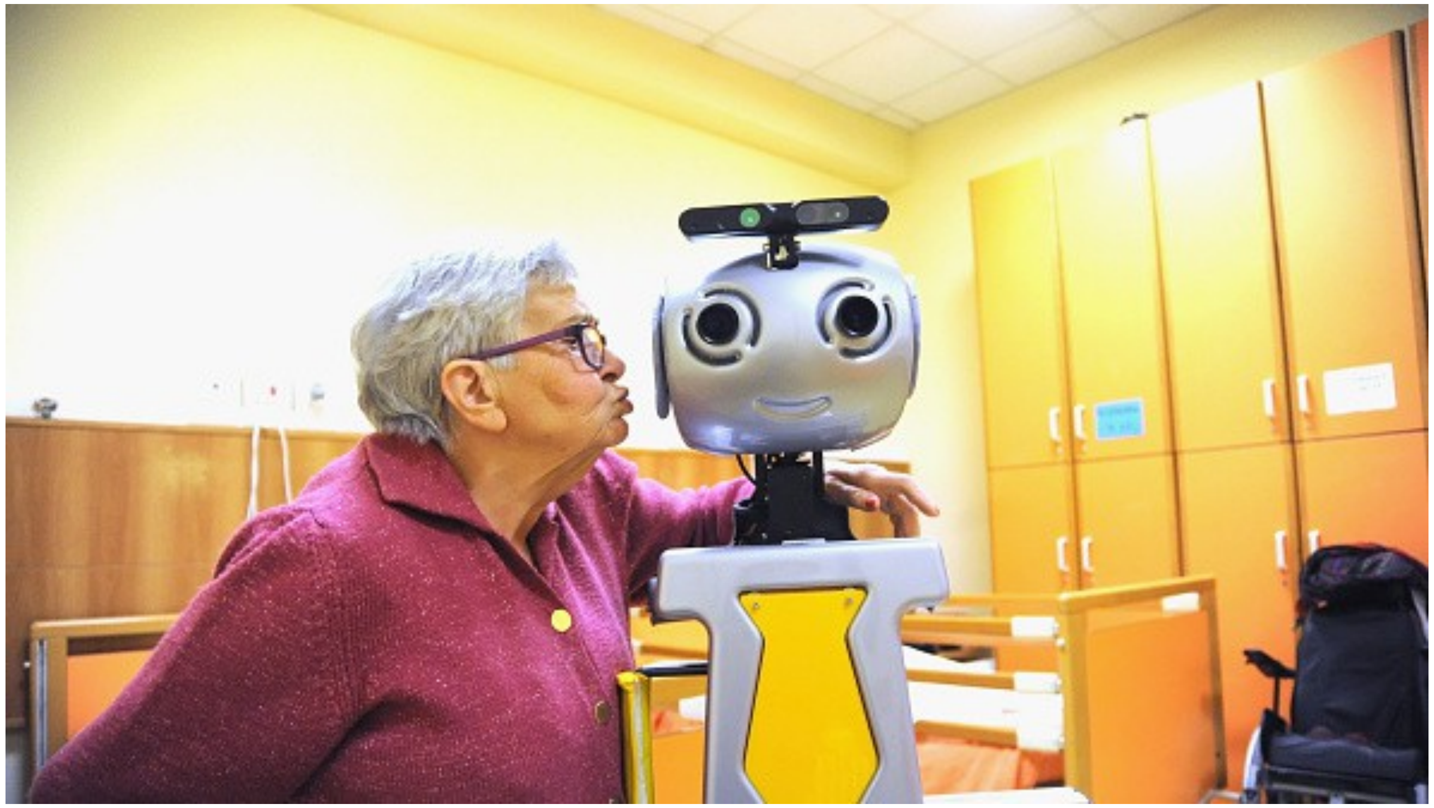
Google Photos, y'all [redacted] up. My friend's not a gorilla.

RETWEETS 3,356
FAVORITES 1,930

8:22 PM - 28 Jun 2015

Human-Robot Interaction

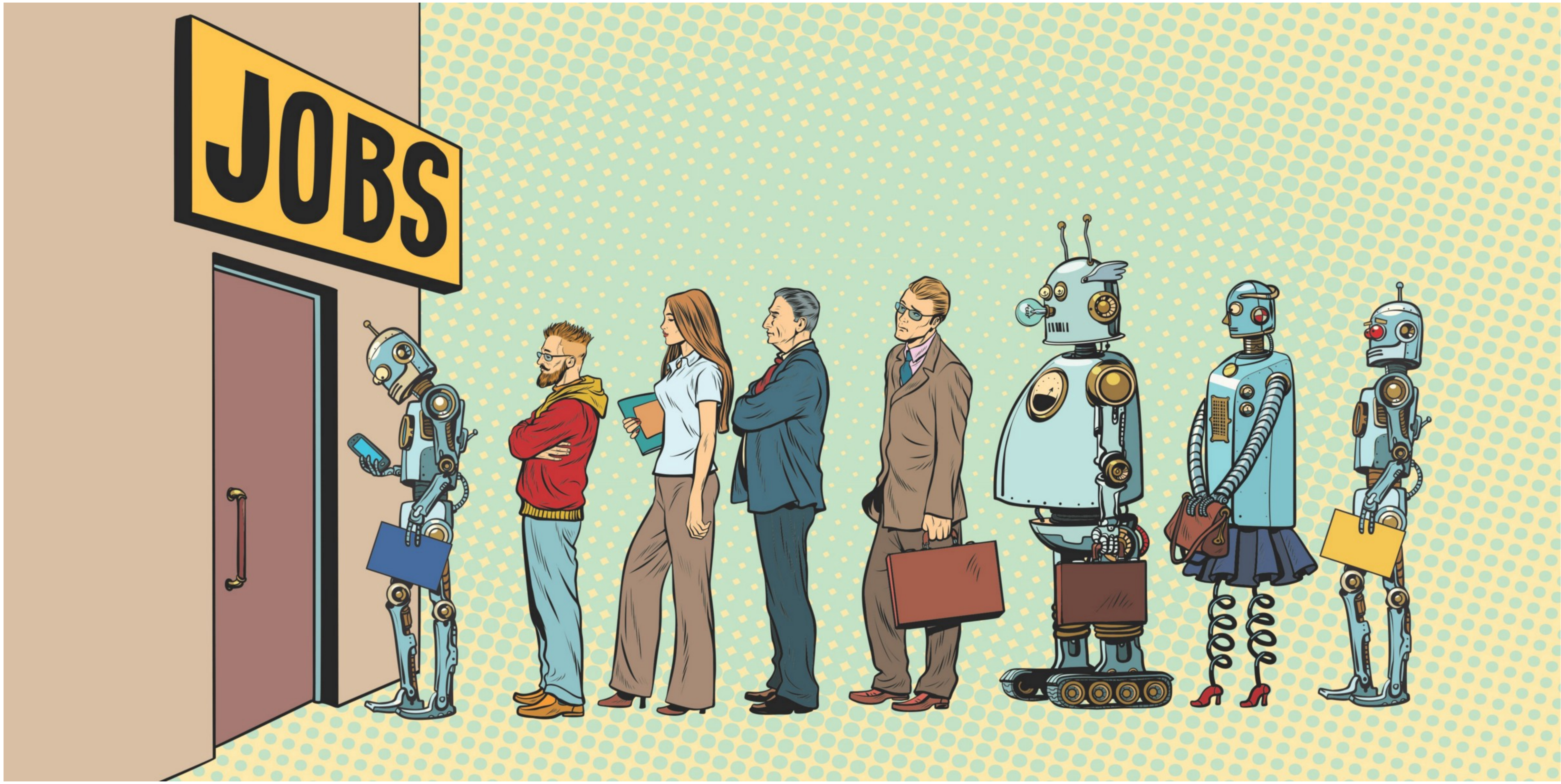




Autonomous systems

smart
On my way

Automation & Employment



- Check wattle on Friday for the first quiz. I will post the quiz by 10am. You have time until Monday 5pm to respond to the question.

- Next week: AI and value alignment — The Problem of Control