Lecture 8-1

How to Make AI Compatible with Human Values?

29-Sep-2020
Dr. Atoosa Kasirzadeh

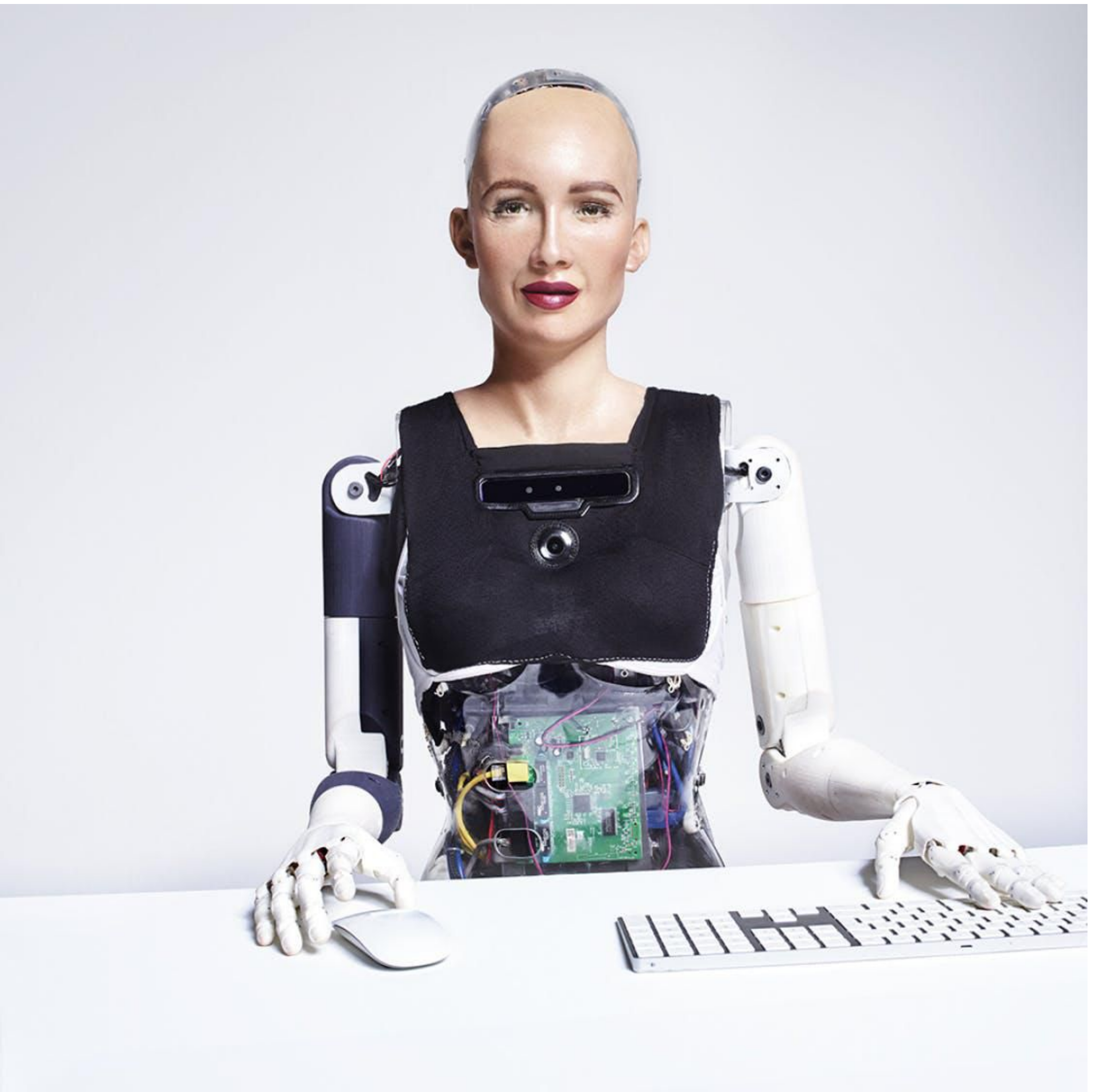
Administrative

- Participation grade: Give a comment in class **<u>OR</u>** in my office hours **<u>OR</u>** post to the Wattle discussion forum (2%) + 4 quizzes (4 x 2%)
  - You should post any interesting news item related to the "ethical and social implications of AI" that you happen to read to the wattle discussion forum, and take a moment to reflect on their significance and relevance to our in-class topics and conversations. Or you could respond to any question in class, engage in a conversation, or come to my office hours with a question/discussion. At least one time engagement (either in class, or in my office hours, or on wattle discussion forum is required. 2%

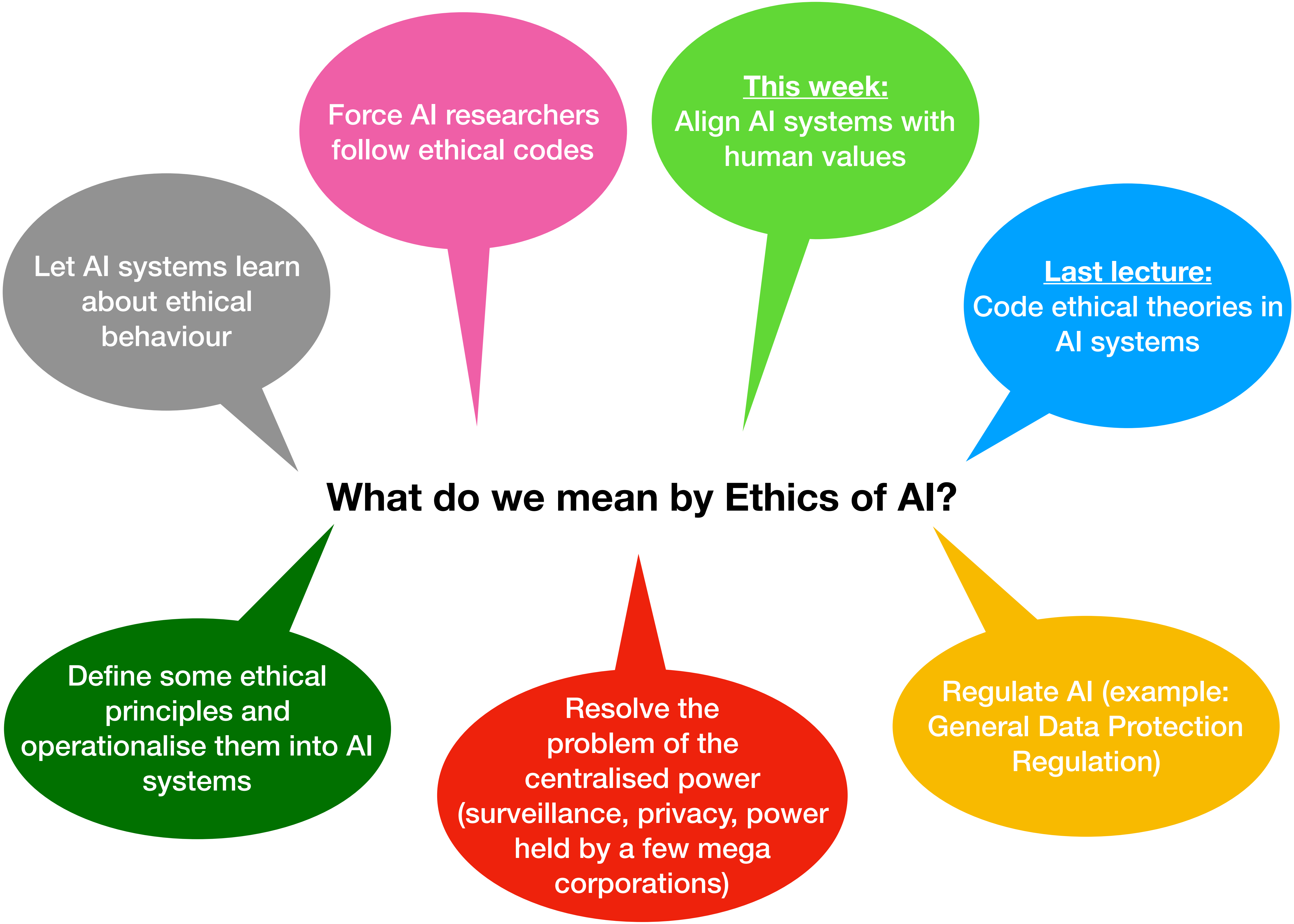
- Zoom Office hours (AEST, every Wednesday from tomorrow, the 30th of September, 4pm—5pm). I will post the link to my virtual zoom meeting room on the announcements today. If you cannot make it to these hours, write to me at atoosa.kasirzadeh@anu.edu.au requesting an appointment.

- Details about the third assignment of the course (the first assignment of the second part) will be discussed tomorrow.

Sophia: the first Humanoid citizen!



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



What do we mean by Ethics of AI?

Let AI systems learn about ethical behaviour

Force AI researchers follow ethical codes

This week:
Align AI systems with human values

Last lecture:
Code ethical theories in AI systems

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)

Regulate AI (example: General Data Protection Regulation)

What do we mean by AI?

(Not mutually exclusive groups)

Origin: 1956 Dartmouth College

1. Logic-based symbolic manipulation (Intelligence with representation)

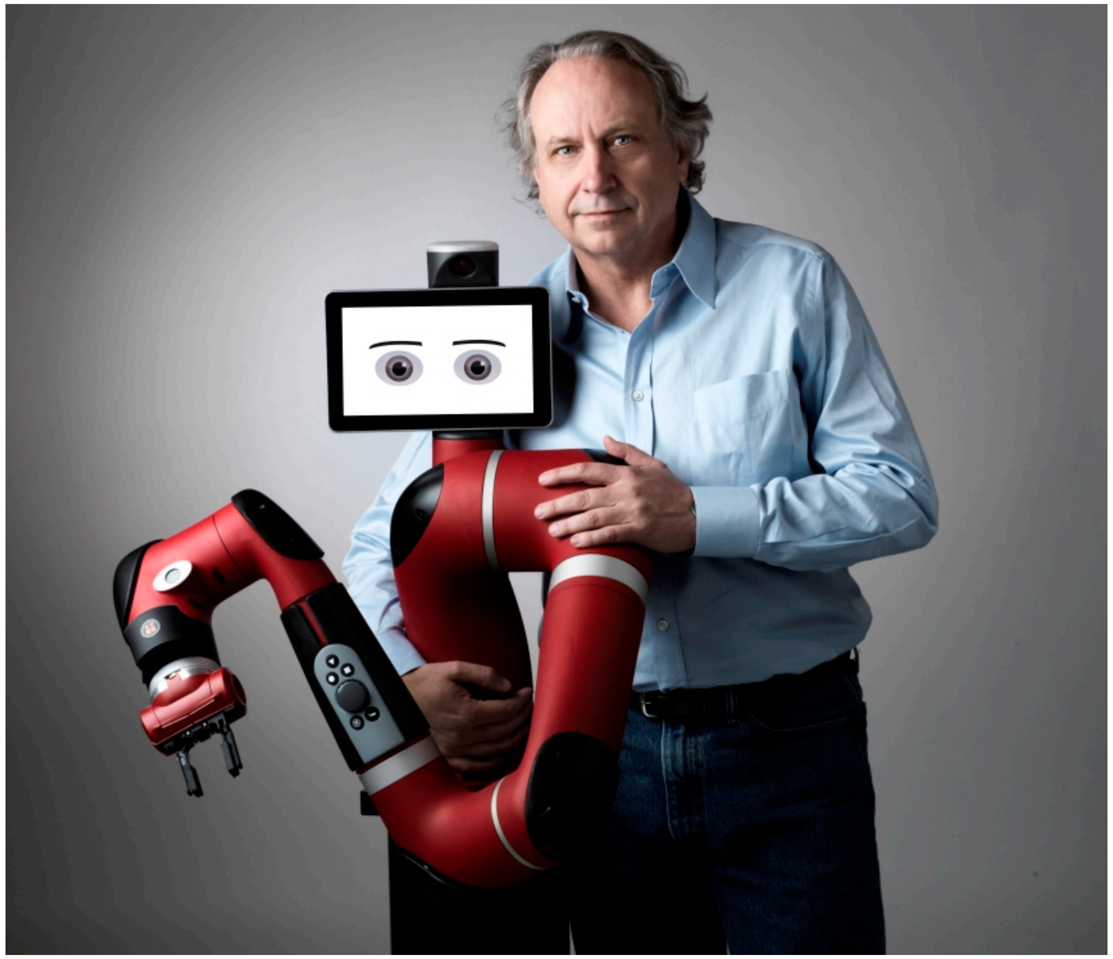
    - The activity of the system as “reasoning” about an external world

    - The reasoning process is implemented by two components

        A. an inference system operating over

        B. a model of the task domain

2. Embodied robotics

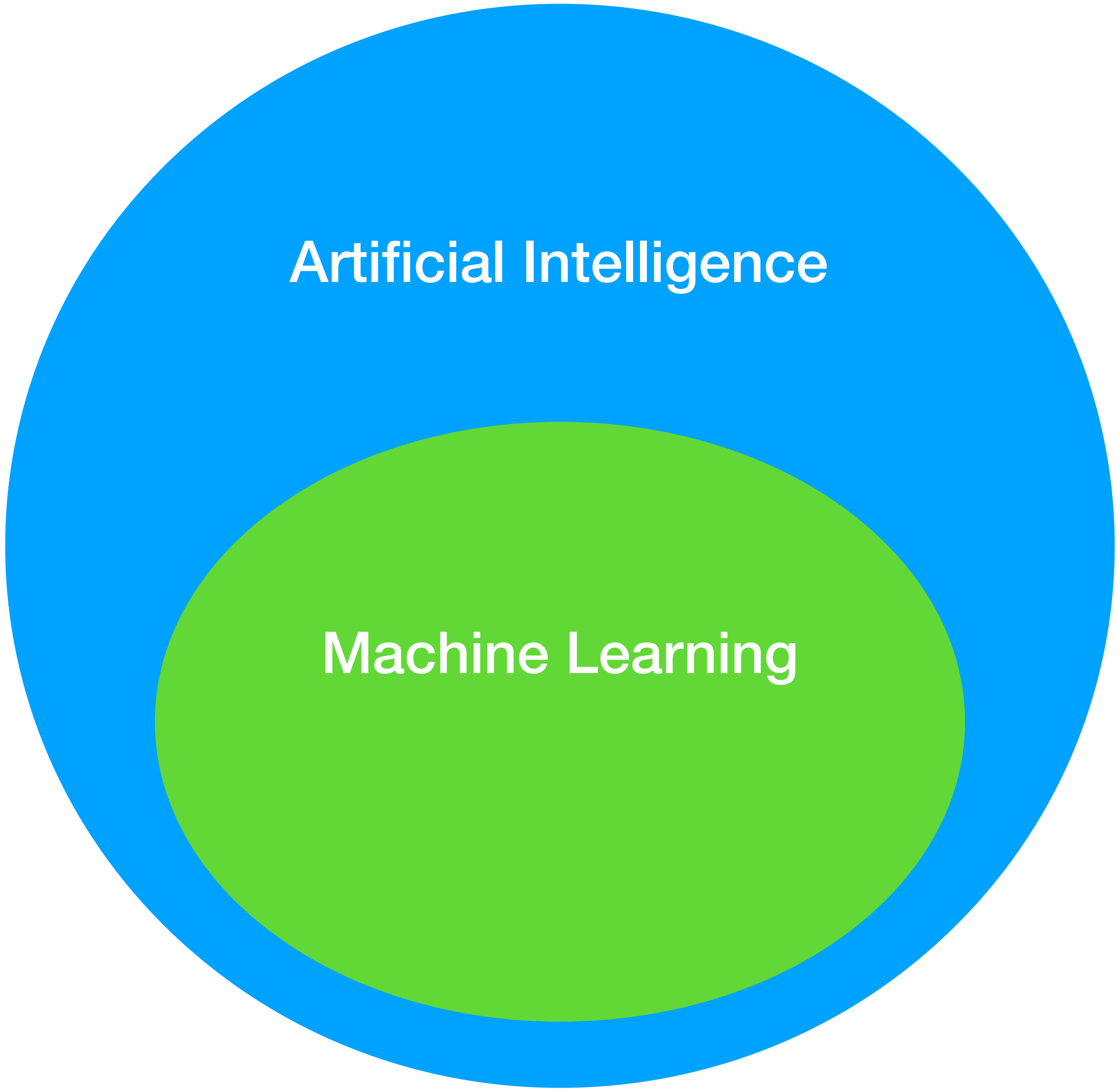


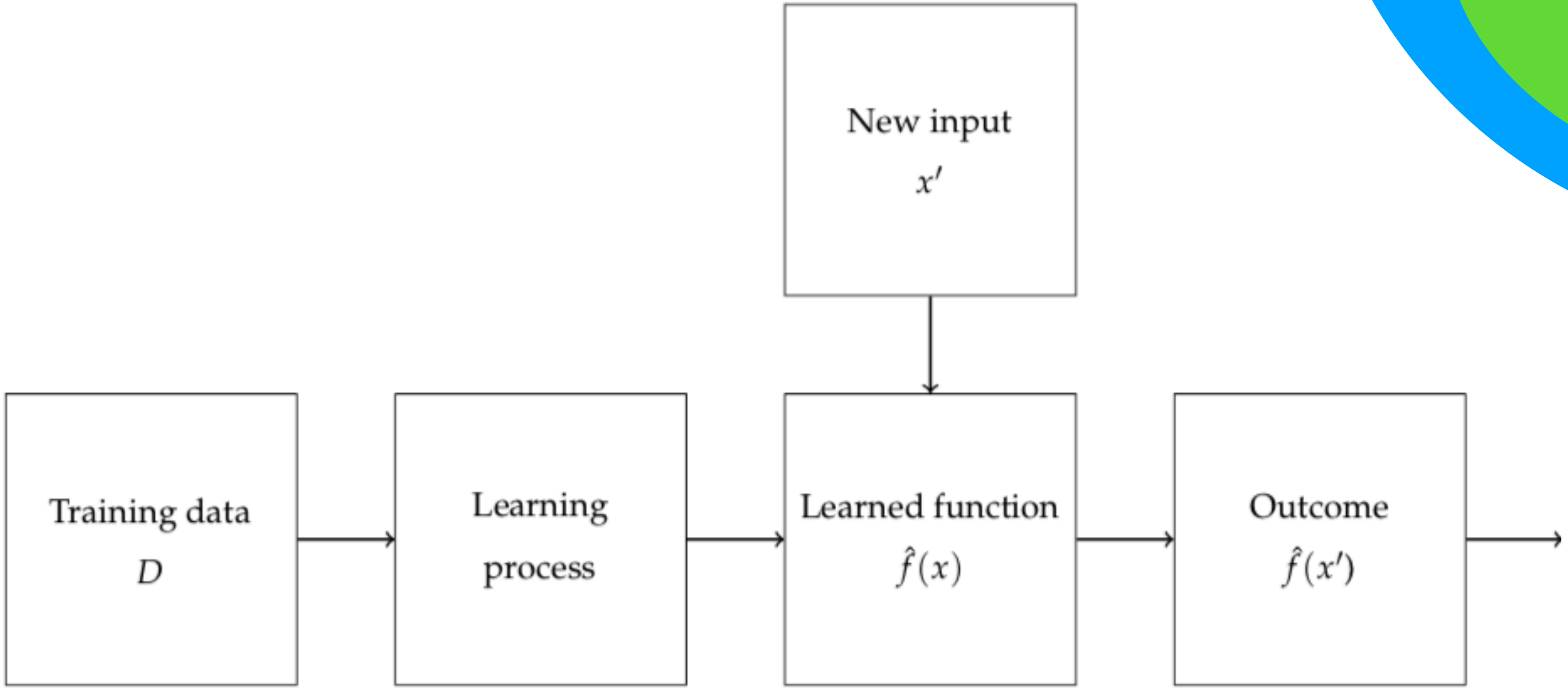
Use the world as its own model; Intelligence without representation

Don’t try to build what we humans think an “intelligent robot” would be like

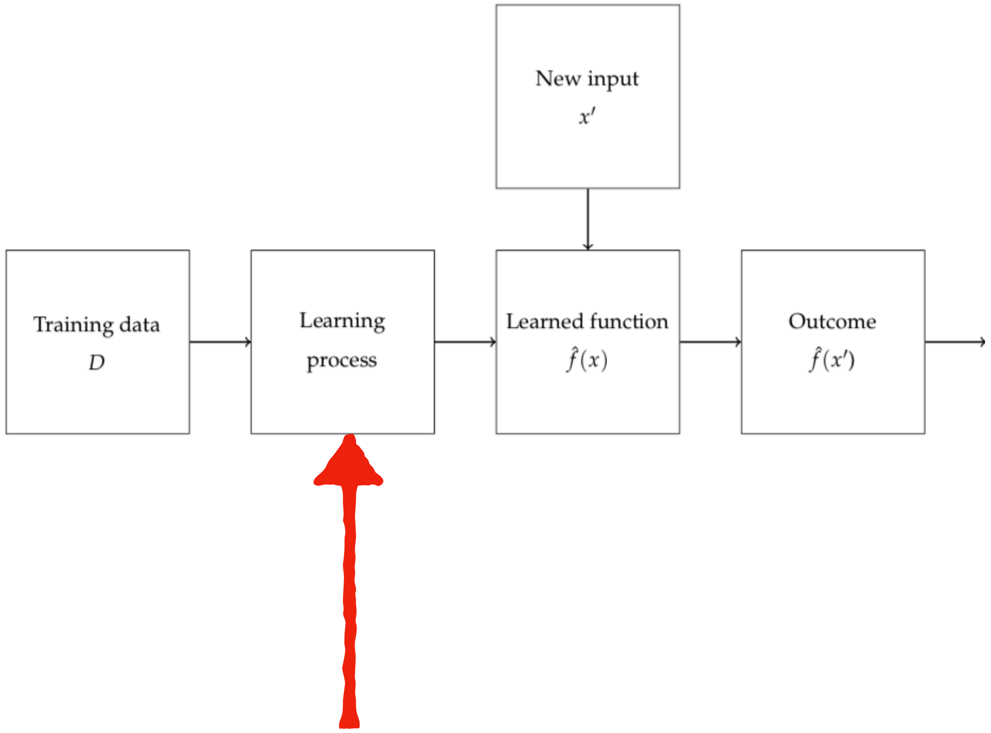
Instead, incrementally build up the capabilities of intelligent systems from the bottom, constructing complete systems at each step of the way and thus automatically ensure that the pieces and their influences are valid

3- Machine learning





Learning process: statistical/optimality



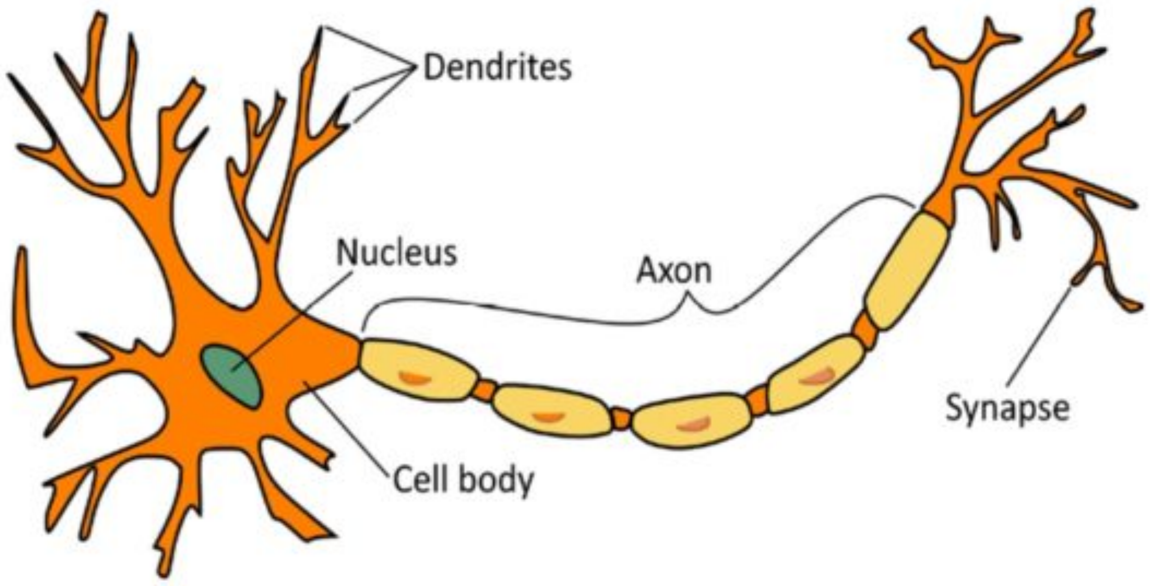
New input x'

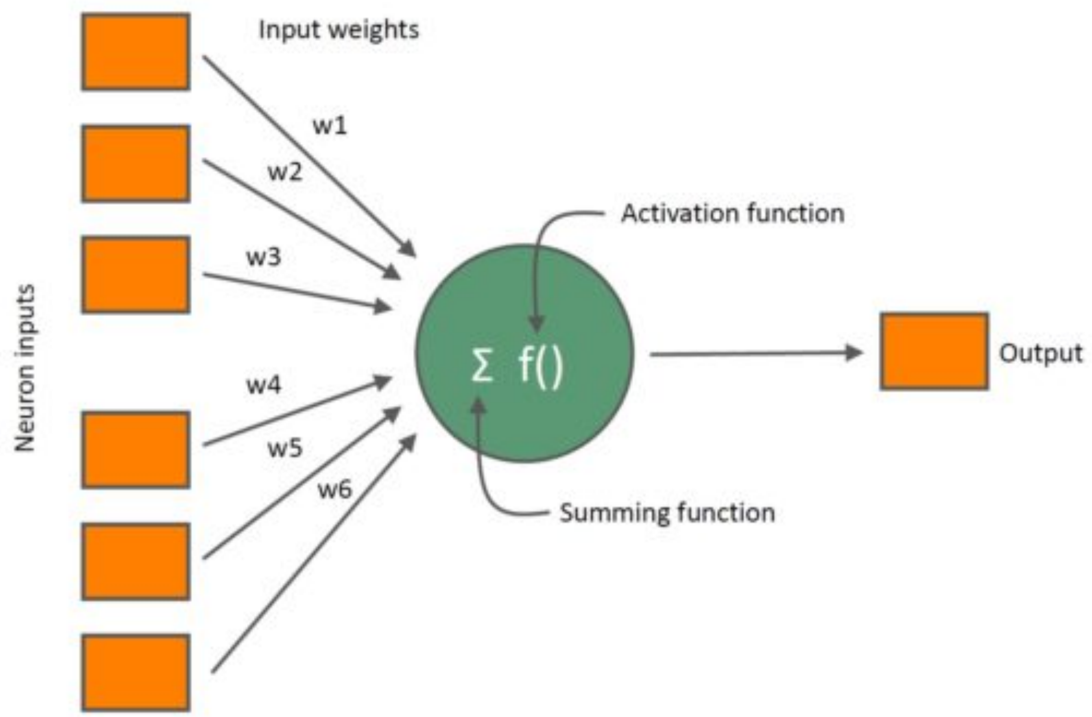
Training data D

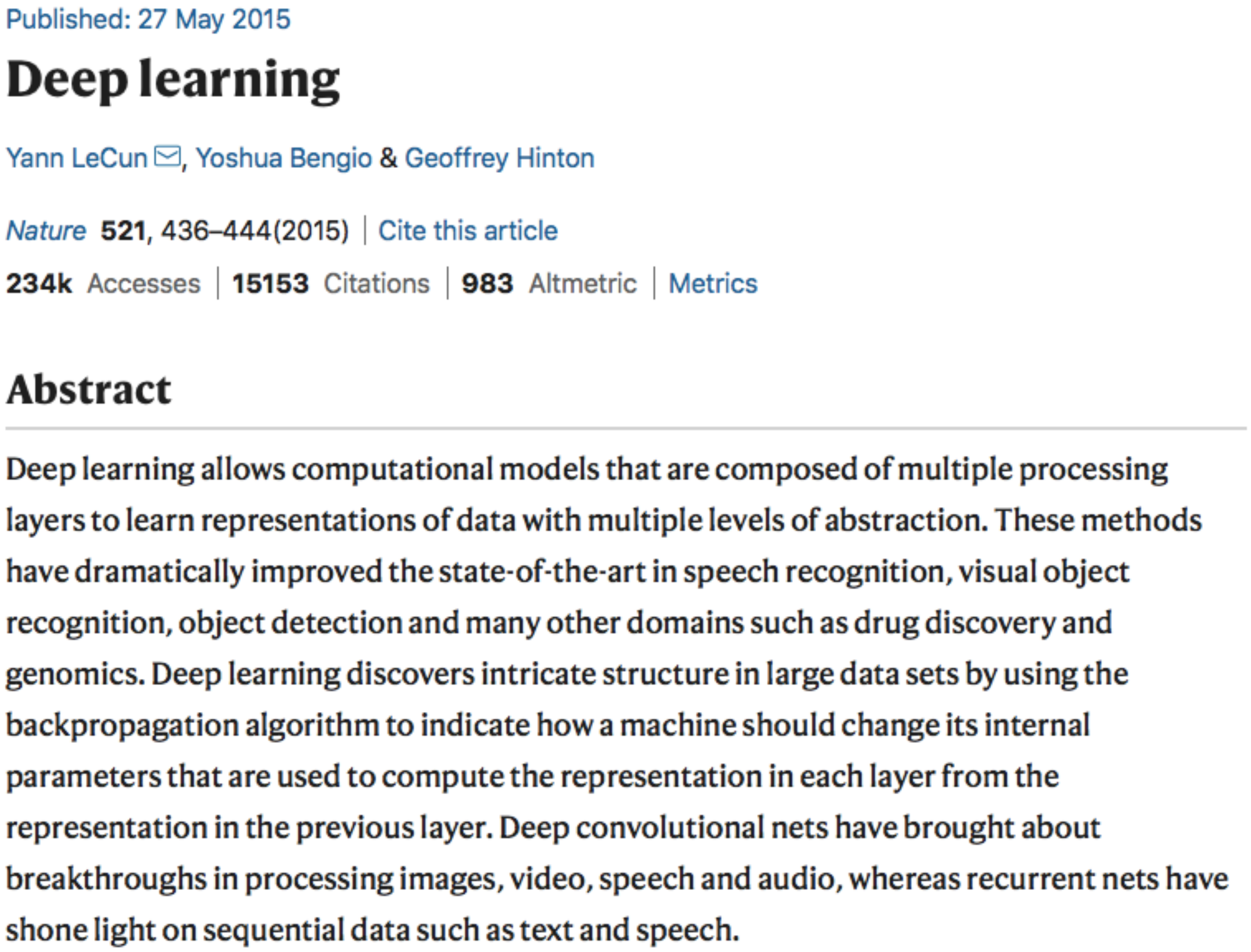
Learning process

Learned function $\hat{f}(x)$

Outcome $\hat{f}(x')$







Published: 27 May 2015

Deep learning

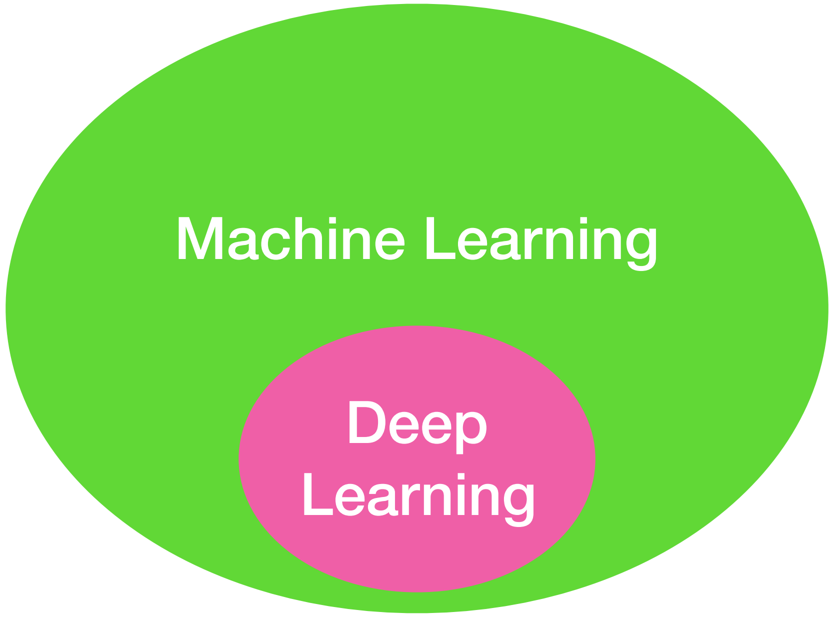
Yann LeCun ✉, Yoshua Bengio & Geoffrey Hinton

Nature **521**, 436–444(2015) | Cite this article

234k Accesses | **15153** Citations | **983** Altmetric | Metrics

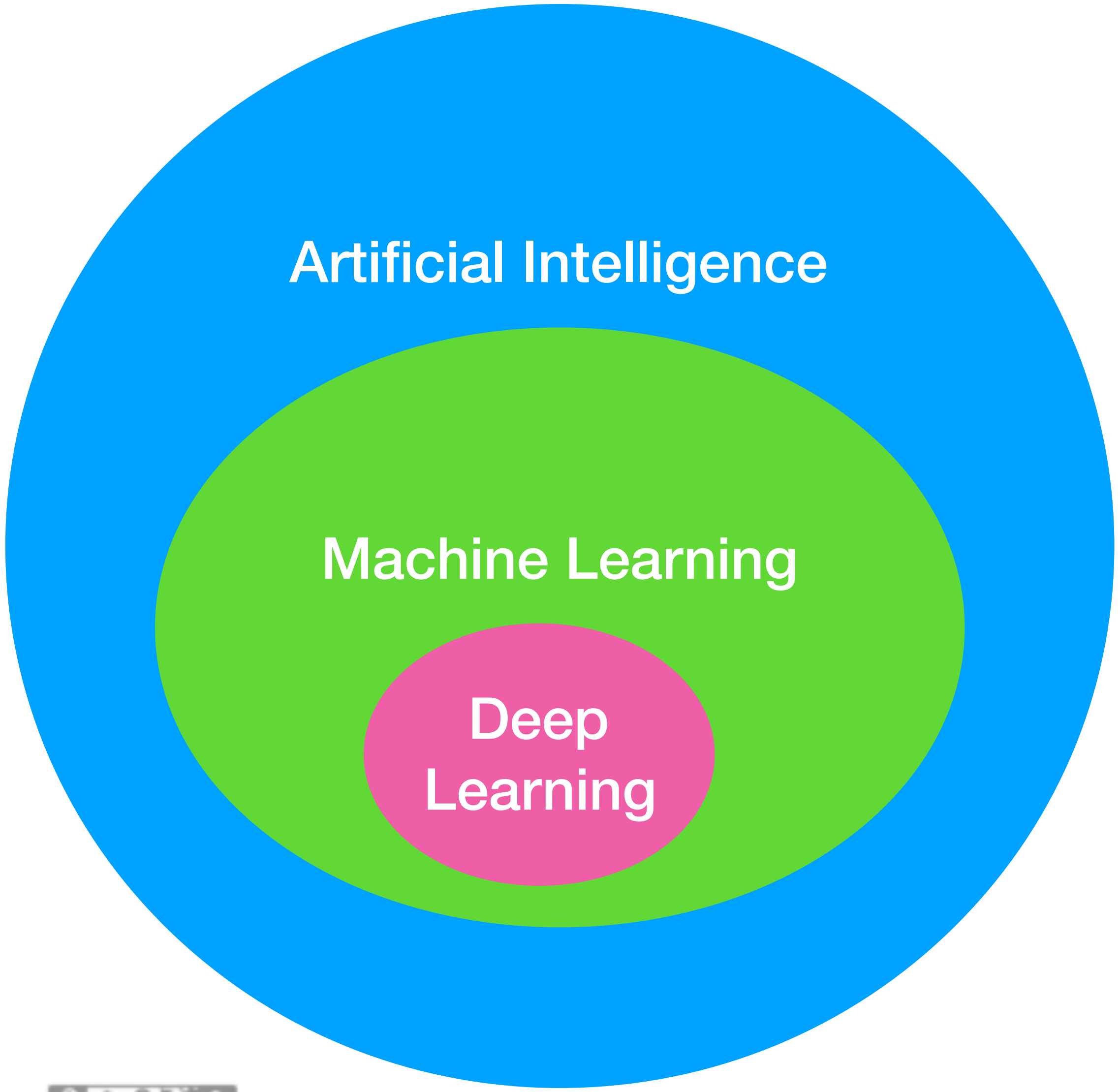
Abstract

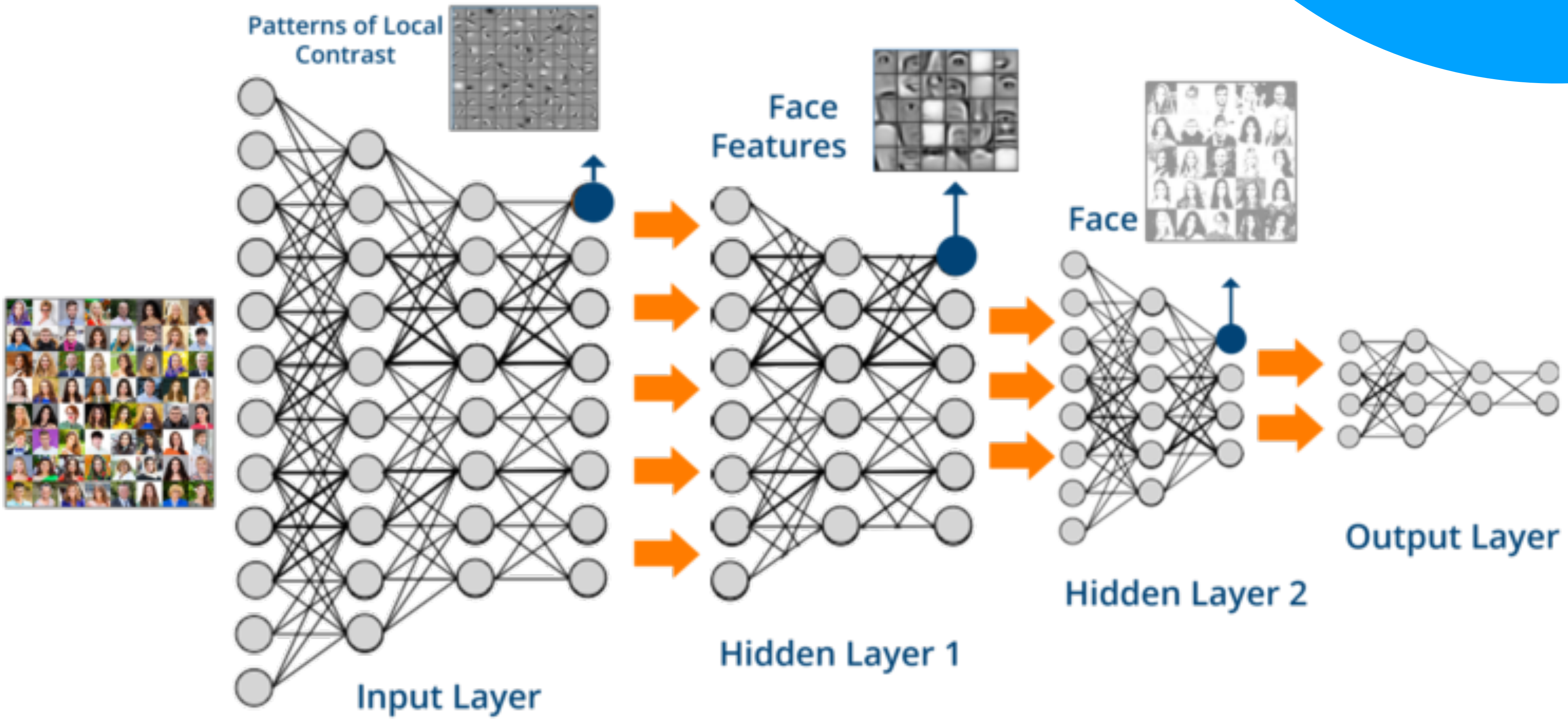
Deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction. These methods have dramatically improved the state-of-the-art in speech recognition, visual object recognition, object detection and many other domains such as drug discovery and genomics. Deep learning discovers intricate structure in large data sets by using the backpropagation algorithm to indicate how a machine should change its internal parameters that are used to compute the representation in each layer from the representation in the previous layer. Deep convolutional nets have brought about breakthroughs in processing images, video, speech and audio, whereas recurrent nets have shone light on sequential data such as text and speech.



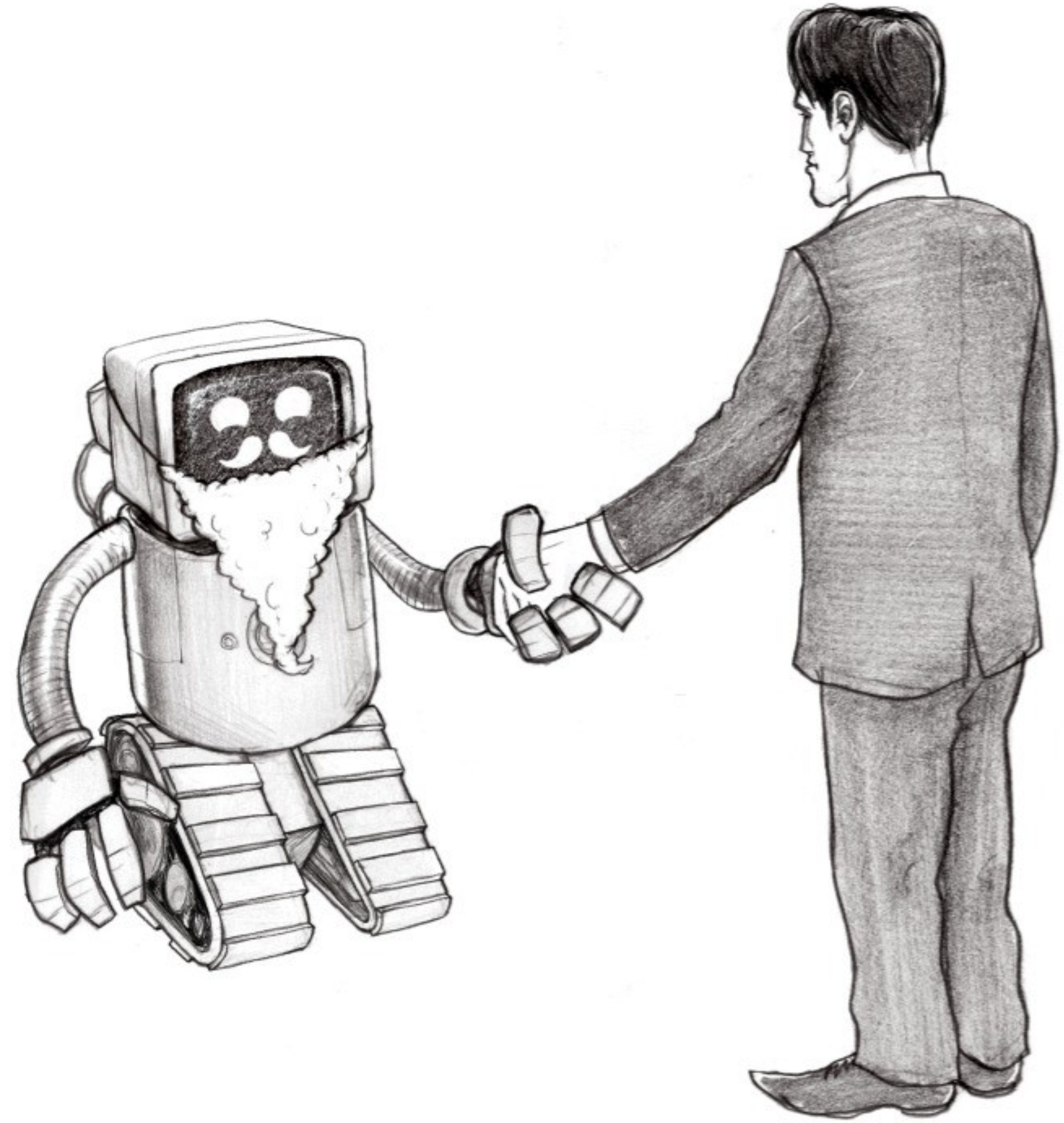
Deep Supervised learning

Image classification





What is AI alignment?

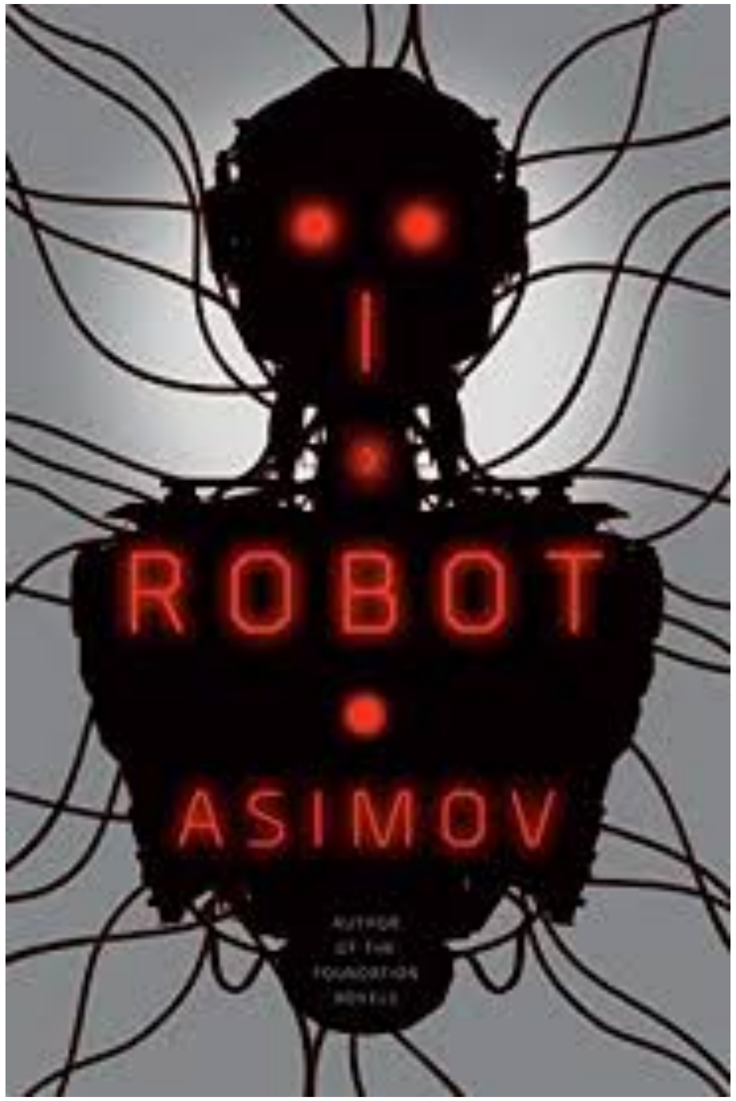


Three Laws of Robotics

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.

2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law

3. A robot must protect its own existence as long as such protection does not conflict with the First or the Second Laws.



On a simple reading of these laws, do you think they can be operationalized into our current AI systems?

What could be some problems with the operationalization of these Laws?