Philosophical Issues in AI

Introduction to Artificial Intelligence

G. Lakemeyer

Winter Term 2016/17

AI from a Philosophical Perspective

Is it really possible to have intelligent machines?
There are two main points of view:

Weak AI: machines which act as if they were intelligent.

$\Rightarrow$Mostly the view of this course.

Strong AI: machines with a real (human like) conscious mind.

Note: Questions like “can machines think?” are often problematic from the outset. E.g., what does it mean to think anyway??

In the following: Some arguments against weak and strong AI.

Weak AI (1)

is impossible because

- machines can only do those things which they are told;
- machines cannot make ethical decision, cannot have emotions, humor, etc.

Weak AI (2)

The mathematical argument:

Gödels incompleteness result:

> For every formal system F which includes arithmetic there is a sentence $G(F)$ with
>
> - $G(F)$ is consistent.
> - $F \nvdash G(F)$
> - $G(F)$ is true in the intended interpretation.

Since computers are formal systems, this is true for them.

But humans (like Gödel) are capable of recognizing the truth of such sentences!

Strong AI—Can machines be conscious?

Turing: Even the question makes no sense. Who knows if other humans are conscious?
... although it seems pretty likely.

Searle's Chinese Room

Description of a thought experiment:

- There is a room with
  - a person who only understands English,
  - a rule book (in English),
  - a stack of paper, some pages are empty; others have strange symbols on them.
- The person receives a piece of paper with strange symbols through a slot in the door.
- The person finds the symbols in the rule book and follows the instructions, in the course of which the person writes strange symbols on the piece of paper.
- In the end the piece of paper is passed back.

From the outside it looks as if both the questions and answers are written in Chinese.
Does the room understand Chinese??

Searle's Argument

Searle's argument against the possibility that machines can be conscious has the following structure:

1. Certain objects do not understand Chinese.
2. The person in the room, the paper, and the rule book are such objects.
3. If each object of a set does not have a conscious mind, then the system composed of these objects cannot be conscious either.
4. Hence the "Chinese Room" does not understand Chinese.

More about Searle's Point of View

Functionalism: a system and all its parts are completely determined by their I/O behavior.

Example:

> Functionalists believe that the way the brain works depends solely on the topology and the I/O behavior of the neurons.
>
> Hence the brain could at least in principle be simulated.

Searle is an anti-functionalist!

His point of view: consciousness is an intrinsic property of the brain.

A problem with this view:

> If consciousness is not determined by the I/O behavior, then it is an epiphenomenon, i.e. it does not have a causal effect.

But then: how can conscious perceptions like pain be explained?

Levesque's Summation Room

Similar room as in the case of Searle, just simpler:

- The input consists of twenty 10-digit number
- The output is a 12-digit number which happens to be sum of the input numbers

What does the rule book look like?

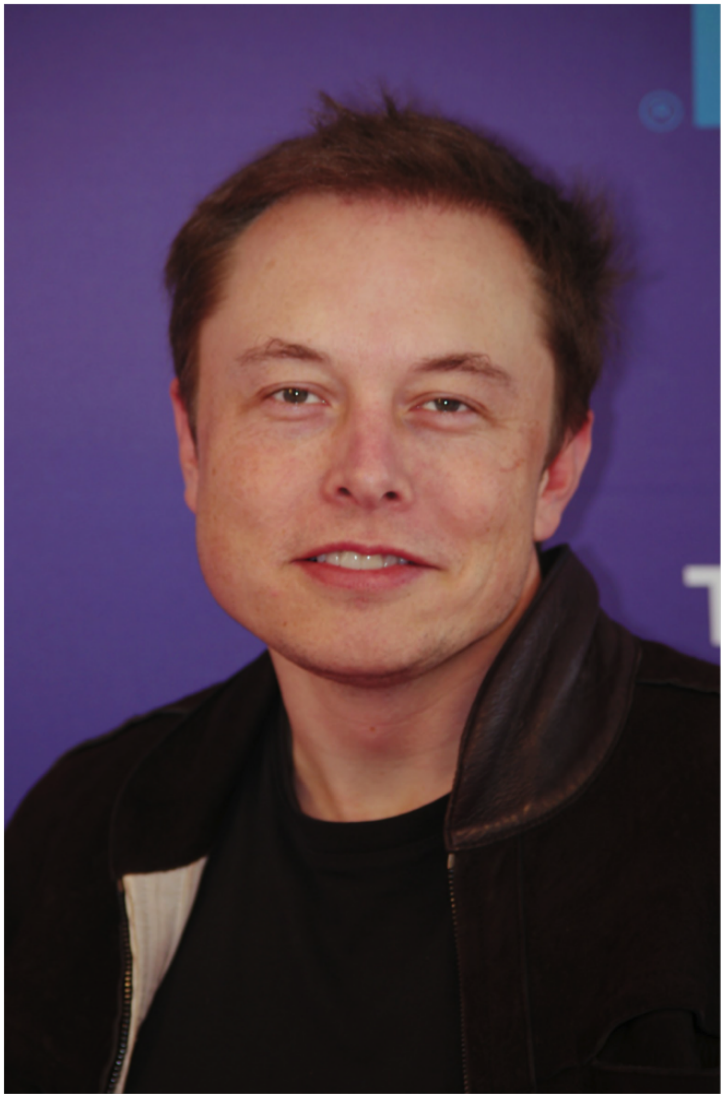
1. If it were a lookup table, then the person would clearly not know how to add (so Searle would be right!)
   But: The lookup table has size 10^{200} and hence is impossible!
2. If it were a description of an algorithm of how to add, then that is certainly feasible, but then we would probably say the person knows how to add!

Steven Hawking

“The development of full artificial intelligence could spell the end of the human race... It would take off on its own, and re-design itself at an ever increasing rate. Humans, who are limited by slow biological evolution, couldn’t compete, and would be superseded.”

Elon Musk

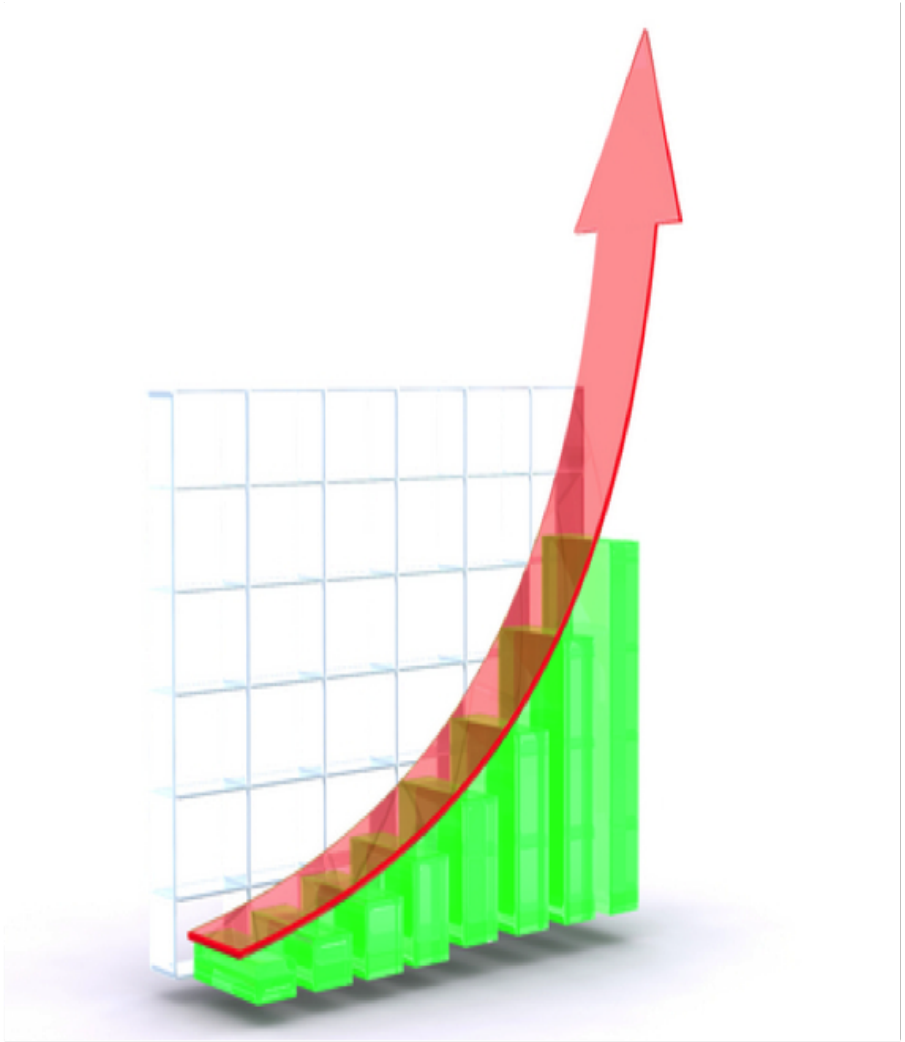
“I think we should be very careful about AI. If I were to guess what our biggest existential threat is, it’s probably that. So we need to be very careful ...With AI we are summoning the demons.”



Technological Singularity

Behind such opinions there is the belief that

- AI will necessarily progress exponentially
- in the end there will be a super intelligence making humans obsolete



Some Arguments against Singularity

- It is true that computers become more and more powerful ("Moore's Law"), but intelligence is much more than quick thinking.
- Progress in software development lags behind hardware development.
- As we know as computer scientists, there are problems which are not solvable even if you are super human.
- Why should human intelligence be a turning point, after which comes super intelligence?
- . . .

AI has made big progress in

- Image and speech recognition
- Question-/Answering Systems (like “Watson”)
- Game Playing (Chess, Poker, Go: http://www.nature.com/nature/journal/v529/n7587/full/nature16961.html)
- Autonomous driving

Some Progress

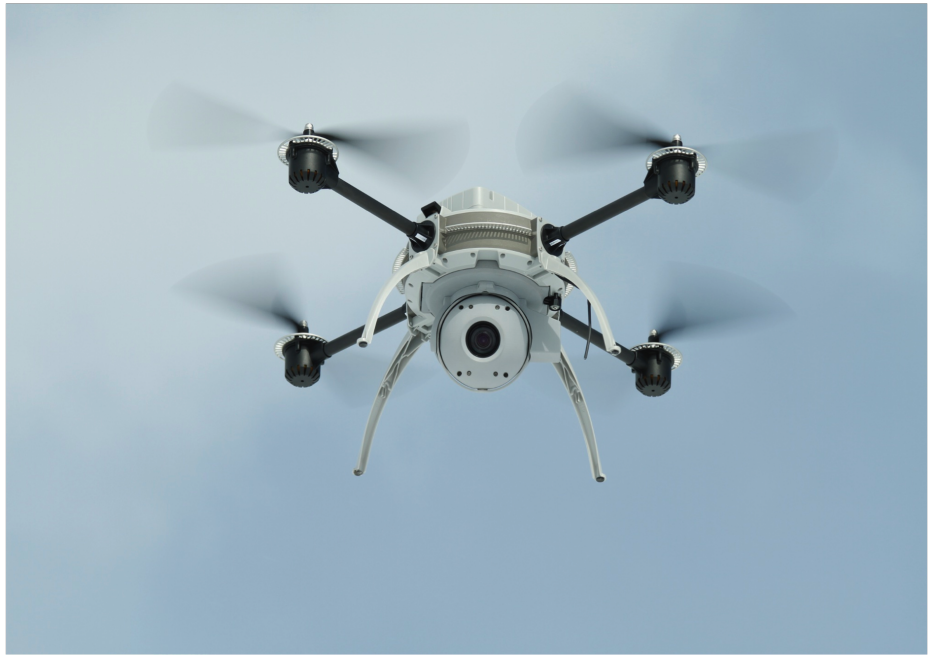
- making rational decisions
- still difficult to react appropriately in unforeseen situations

Little Progress

- Ethical decisions.
- Emotions
- Intuition
- Consciousness

In AI we are not really worried about. . .

. . . but perhaps we should be worried about . . .



- It is possible to build autonomous systems.
- They may be cheap.
- They can be used to harm people.
- They may fall into the wrong hands.
- They are only (sigh!) moderately intelligent.
- . . .