Philosophy of AI (WBMV05003)

Janneke van Lith


Universiteit Utrecht

2019-2020 period 3

February 5, 2020

Outline

Philosophical issues w.r.t. AI

- Philosophy of mind

  How do mind and matter relate? Can machines think? What is intelligence? What is it to have thoughts *about* something? How do internal states function as representations? What is it to have conscious beliefs, feelings, etc.? Could machines be conscious?

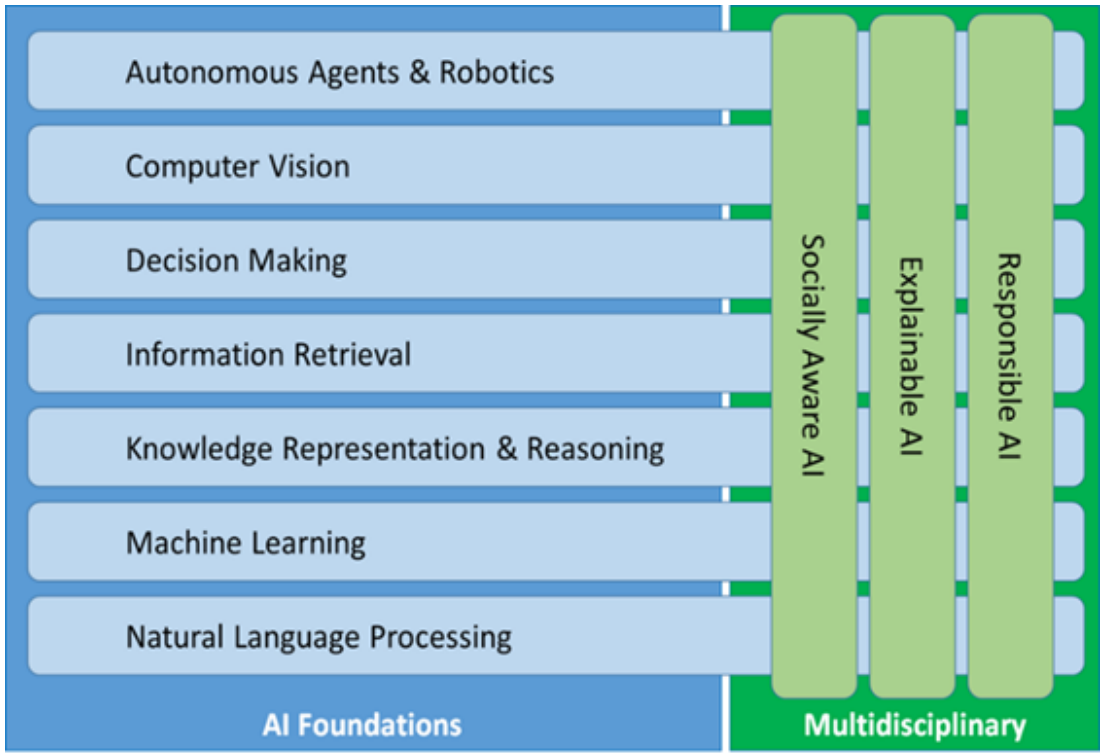
- Methodology and Philosophy of science

  How to do AI? Which approaches, subfields, architectures are there? How does AI relate to other fields?

- Ethics

  What is the ethical impact of AI (on its users, geo-politics, economy, culture . . .)? How to do AI ethically? How to formalize and implement responsibility? Can machines be moral agents?

Dutch AI manifesto



Goals of this course

- Learn about fundamental issues in philosophy of AI
- Learn about current discussions in philosophy of AI
- Practice argumentation and presentation skills
- (and perhaps get inspiration for a Master's Thesis on a topic related to this course)

Course setup

Three parts of the course:

- Introductory lectures, entrance test
- Five themes, with a lecture and a seminar each:
  1. What is a good test for intelligence?
  2. Ethics and AI
  3. Decision making
  4. AI and Data Science
  5. Responsibilty in intelligent systems
- Final paper

- All course info is on Blackboard; please check the syllabus
- If you need to switch seminar groups, please contact the Student Desk GW today

The Cambridge handbook

Material for the Entrance test: Ch. 1-6

1. History, motivations, and core themes (Franklin)
2. Philosophical foundations (Arkoudas & Bringsjord)
3. Philosophical challenges (Robinson)
4. GOFAI (Boden)
5. Connectionism and neural networks (Sun)
6. Dynamical systems and embedded cognition (Beer)

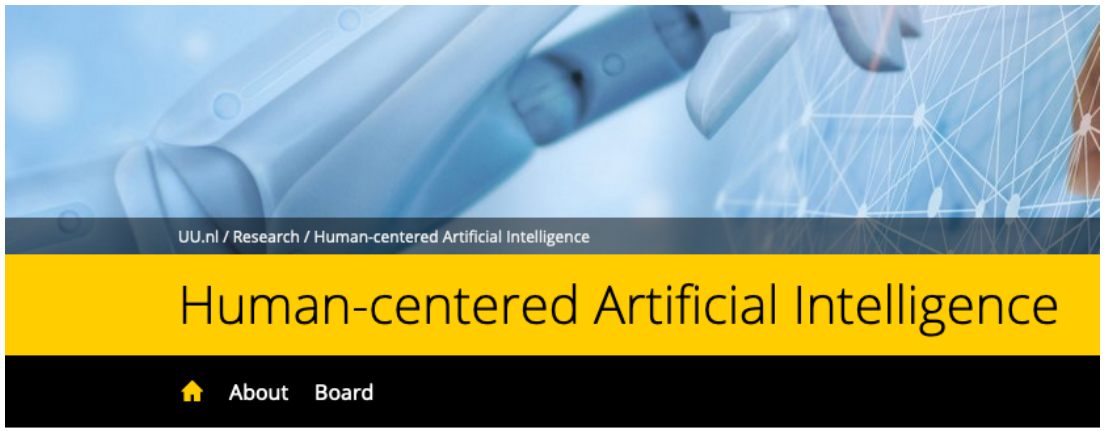
Overview

Artificial Intelligence

Artificial Intelligence is the science of making machines do things that would require intelligence if done by men. – Marvin Minsky

- Smart software vs. Cognitive modelling (Ch. 1)
- Weak AI vs. Strong AI

Artificial Intelligence



Artificial intelligence is the discipline that pursues the understanding, artificial replication and possible enhancement of human intelligence.

The focus area Human-centered
activities undertaken at the depa
University. AI in Utrecht has a ur
departments including compute

Artificial intelligence is the disci
replication and possible enhance
main method is to construe philo
computational models for a larg
such as reasoning and argument
perception, learning, planning, t

artificial intelligence contribute to the scientific understanding of all these phe

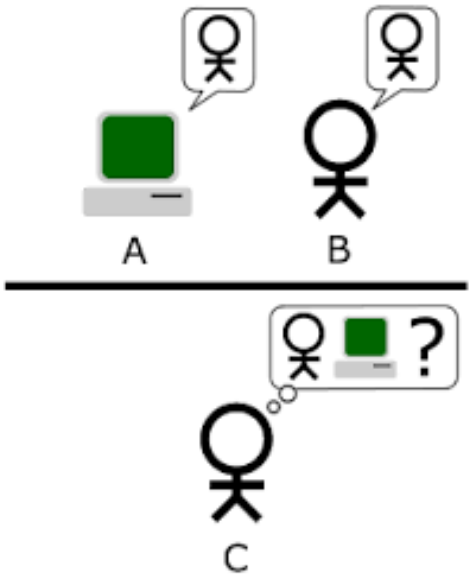
Weak AI thesis

- The computer is (only) a powerful aiding tool for the study of the human mind.
- It is possible to construct machines that perform useful "intelligent" tasks assisting human users.

Strong AI thesis

- An adequately programmed computer has a cognitive state; computer programs explain human cognition.
- It is possible to devise machines that behave like people and possess human capabilities, such as the ability to think, reason, . . . , play chess, walk, . . . , have emotions, pain, be creative, . . .

The Turing Test

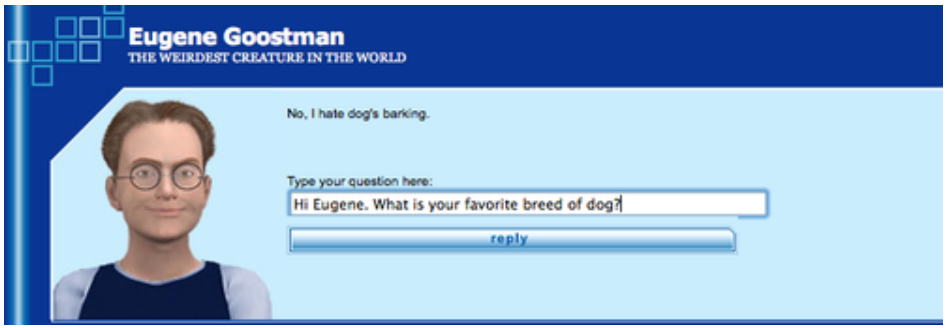


'I propose to consider the question, "Can machines think?" . . . I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words. The new form of the problem can be described in terms of a game which we call the 'imitation game'.' (Turing 1950)

Has the TT already been passed?

- ELIZA
- PARRY
- Eugene Goostman

Yearly Loebner Prize contest (since 1991); the organizer, Kevin Warwick, has claimed that Eugene Goostman has passed the TT in the 2014 edition of the contest.



Assessing the TT

Is the Turing Test a good test for intelligence?

The test is too strong:

- Chimpanzee objection: animals and small children will not pass the TT

The test is too weak:

- Sense organ objection: the TT only tests verbal behaviour; it ignores meaning and causal relations with the world outside the machine.
- Simulation objection: a computer program that only simulates intelligence can pass the TT.
- Blockhead objection: the TT only tests outward behaviour; it ignores the inner workings of the machine.

Assessing the TT

The Blockhead objection emphasises the behaviouristic nature of the TT.

- How do we assess whether other people think / are intelligent? Don't we conclude this on the basis of their outward behaviour?
- A behaviouristic test ignores internal mechanisms, composition and functional organization, but that's a good thing: there might be intelligence based on a substrate that radically differs from ours.

Still, most AI researchers would agree that the objections establish the invalidity of the TT.

How to improve on the TT

- output criterion

Computers must be indistinguishable from the human not only in conversation, but also in the performance of other tasks. The Lovelace Test is aimed at creativity in making artifacts. Harnad's Total Turing Test includes sensorimotor capability.

- design criterion

The internal workings of the machine should somehow count. (Perhaps the program should be modular, or human-like.)

How to characterize intelligence

Copeland (1993) identifies thinking with massive adaptability:

> *"[T]he concept of a thinking thing serves to mark out a distinction between, on the one hand, those entities whose wealth of action-directing inner processes renders them inventive, expansively adaptable, capricious agents, and on the other hand those entities with a paucity of such processes, whose behaviour is correspondingly rigid, circumscribed, and stereotyped."*

Robinson (Handbook, Ch. 3) introduces the similar notion of flexibility, i.e. of being able to respond appropriately to a wide range of novel circumstances.

Searle requires even more of intelligence, as we shall see (understanding, intentionality).

Overview

Are we computers?

Possible objections to machine intelligence:

- People have minds, in contrast with computers.
- People have free will, in contrast with computers.
- People are conscious beings / have conscious states, in contrast with computers.
- People have original intentionality, in contrast with computers.

Therefore, we aren't computers, and machines can't be as intelligent as people.

The ontology of the mind

- Dualism
- Monism
  - Materialism
    - reductive materialism
    - eliminative materialism
  - Idealism
- Functionalism

Dualism

The essential nature of conscious intelligence resides in something non-physical.

- Substance dualism holds that each mind is a distinct non-physical thing, made of a non-physical substance.
- Property dualism holds that although there is no special mental substance, the brain has a special set of properties that are non-physical.

Figure: René Descartes (1596-1650)

Materialism

Only matter exists. Mental states / processes are identical to physical states / processes (type vs. token).

Materialism

Only matter exists. Mental states / processes are identical to physical states / processes (type vs. token).

How do the sciences of matter (s.a. neuroscience) and of mind (s.a. intentional psychology) relate?

- reductive materialism: (future versions of) neuroscience will reduce to (future versions of) intentional psychology. Compare intertheoretic reductions in the physical sciences: light is electromagnetic radiation; temperature is mean kinetic energy.
- eliminative materialism: No one-to-one match-ups between mental and physical states will be found, because our common-sense psychological framework is false and radically misleading. "Folk psychology" should be rejected.

Functionalism

According to functionalism, the defining feature of a mental state is its functional / causal relation to the environment, to other mental states, and to bodily behaviour. So, the same mental state could be realised in many different ways, and by different kinds of systems (brains, computers).

Example: what makes something a state of pain, depends not on its internal constitution, but on whether it is typically caused by bodily injury, leads to the desire to get out of that state, etc.

Freedom: the Free Will problem

(from: SEP entry on Compatibilism)

1. Some person (qua agent), at some time, could have acted otherwise than she did.
2. Actions are events.
3. Every event has a cause.
4. If an event is caused, then it is causally determined.
5. If an event is an act that is causally determined, then the agent of the act could not have acted otherwise than in the way that she did.

This leads to a conflict. How to resolve it?

Freedom: computers and humans

Hard determinism: deny 1.

neurophysiological determinism implies that humans have no free will; therefore, neither computers nor humans have free will; so there is no fundamental difference between humans and machines.

Compatibilism: deny 5.

humans and machines can have free will; so there is no fundamental difference between humans and machines. Free will and determinism are compatible.

Freedom: the Consequence argument

Consequence argument. If determinism is true, then my actions are determined by laws of nature and events prior to my birth. These are not up to me; so neither are their consequences.

This is a problem for compatibilism.

Consciousness

Three different notions of consciousness:

- base notion: being awake, experiencing;
- internal monitor: being able to think about one's own cognitive states and processes;
- qualia: the sensory episodes have "feely properties".

Especially the third poses problems for the possibility of machine consciousness. (Famous thought experiments by Frank Jackson and Thomas Nagel.)

Another useful distinction: state consciousness vs. creature consciousness. More information: SEP.

Frank Jackson: What Mary doesn't know

Imagine a scientist Mary. She knows every physical fact there is to know, but she never leaves her black-and-white room. One day she finally leaves and sees a red tomato. At that point she learns something new (or so Jackson claims). This shows that complete knowledge of the physical universe still leaves out facts of subjective experience.

Overview

Challenges to Strong AI

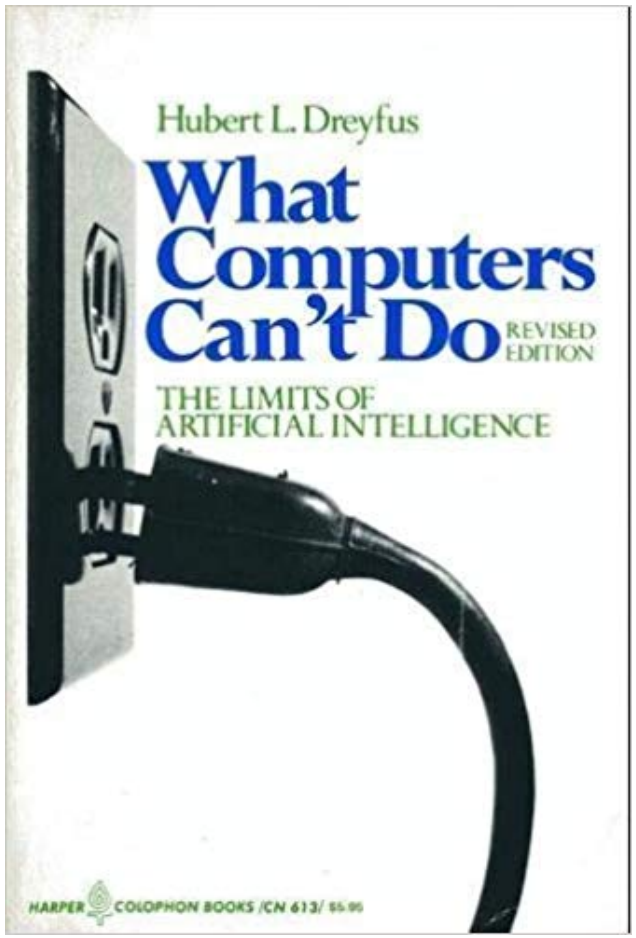
- Heideggerian arguments (Dreyfus; see Ch. 2, 3, 4 and 6)
- The China Brain Argument (Block; see Ch. 2)
- The Chinese Room Argument (Searle; see Ch. 2, 3 and 4)
- Gödelian arguments (Lucas, Penrose; see Ch. 3)

Note: some of these arguments are aimed specifically at GOFAI.

Dreyfus' arguments against AI

Inspired by continental philosophy (Heidegger, Merleau-Ponty)

- rules
- relevance
- know-how
- situatedness



The China Brain Argument

- Thought experiment by Ned Block (1978)
- Imagine the people of China to simulate the workings of the neurons in a brain, communicating by radio. Would such a China Brain be conscious / have mental content? Intuitively: no!
- Aimed specifically at machine functionalism, which (according to Block) would have to say yes.