Philosophy of AI

Chapters 26–27, Russel & Norvig

Peter Ljunglöf
19th April, 2013

What is AI?

Weak AI – acting intelligently

- the belief that machines can be made to act as if they are intelligent

Strong AI – being intelligent

- the belief that those machines are actually thinking

Most AI researchers don't care

- "the question of whether machines can think… is about as relevant as the question of whether submaries can swim" (Dijkstra, 1984)

Is weak AI possible?

There are different opinions…

- …some are slightly positive:
  - ”every […] feature of intelligence can be so precisely described that a machine can be made to simulate it” (McCarthy et al, 1955)
- …and some lean towards the negative:
  - ”AI […] stands not even a ghost of a chance of producing durable results” (Sayre, 1993)

It’s all in the definitions:

- what do we mean by ”thinking” and ”intelligence”?

Alan Turing

The most important paper in AI, of all times:

- (…I'm not the only one who thinks that…)
- "Computing Machinery and Intelligence" (Turing, 1950)
  - introduced the "imitation game" (Turing test)
  - discussed objections against intelligent machines, including almost every objection that has been raised since then
  - it's also easy to read…

    …so you really have to read it!

Turing's objections [1–3]

(1) The Theological Objection

- "Thinking is a function of man's immortal soul. God has given an immortal soul to every man and woman, but not to any other animal or to machines. Hence no animal or machine can think."

(2) The "Heads in the Sand" Objection

- "The consequences of machines thinking would be too dreadful. Let us hope and believe that they cannot do so."

(3) The Mathematical Objection

- Based on Gödel's incompleteness theorem.

Turing's objections [4–5]

(4) The Argument from Consciousness

- "No mechanism could feel […] pleasure at its successes, grief when its valves fuse, […], be angry or depressed when it cannot get what it wants."

(5) Arguments from Various Disabilities

- "you can make machines do all the things you have mentioned but you will never be able to make one to do X."
- where X can… "be kind, resourceful, beautiful, friendly, […], have a sense of humour, tell right from wrong, make mistakes, fall in love, enjoy strawberries and cream, […], use words properly, be the subject of its own thought, […], do something really new."

Turing's objections [6–8]

(6) Lady Lovelace's Objection

- "The Analytical Engine has no pretensions to originate anything. It can do whatever we know how to order it to perform"

(7) Argument from Continuity in the Nervous System

- "one cannot expect to be able to mimic the behaviour of the nervous system with a discrete-state system."

(8) The Argument from Informality of Behaviour

- "if each man had a definite set of rules of conduct by which he regulated his life he would be no better than a machine. But there are no such rules, so men cannot be machines."

The final objection [9]

(9) The Argument from Extrasensory Perception

- this was the strongest argument according to Turing…
- ”the statistical evidence […] is overwhelming”
- ”Let us play the imitation game, using as witnesses a man who is good as a telepathic receiver, and a digital computer. The interrogator can ask such questions as 'What suit does the card in my right hand belong to?' The man by telepathy or clairvoyance gives the right answer 130 times out of 400 cards. The machine can only guess at random, and perhaps gets 104 right, so the interrogator makes the right identification.”

Strong AI

The brain replacement experiment

- by Searle (1980) and Moravec (1988)
- suppose we gradually replace each neuron in your head with an electronic copy
  - what will happen to your mind, your consciousness?
  - Searle thinks that you will gradually feel dislocated from your body
  - Moravec thinks you won't notice anything

Strong AI

The Chinese room experiment (Searle, 1980)

- an English-speaking person takes input and generates answers in Chinese
  - he/she has a rule book, and stacks of paper
  - the person gets input, follows the rules and produces output
- i.e., the person is the CPU, the rule book is the program and the papers is the storage device

Does the system understand Chinese?

The technological singularity

Will AI lead to a superintelligence?

- "…ever accelerating progress of technology and changes in the mode of human life, which gives the appearance of approaching some essential singularity in the history of the race beyond which human affairs, as we know them, could not continue" (von Neumann, mid-1950s)
- "We will successfully reverse-engineer the human brain by the mid-2020s. By the end of that decade, computers will be capable of human-level intelligence." (Kurzweil, 2011)
- "There is not the slightest reason to believe in a coming singularity." (Pinker, 2008)

Ethical issues of AI

Possible risks

- AI might be used towards undesirable ends
  - e.g., surveillance by speech recognition, detection of ”terrorist phrases”
- AI might result in a loss of accountability
  - what's the legal status of a self-driving car?
  - or a medical expert system?
- AI might mean the end of the human race
  - what if the new superintelligent race won't obey Asimov's robot laws?

Peter Norvig's view on the history and future of AI

Can we predict the future of AI?

- "The History and Future of Technological Change"
- a talk by Peter Norvig, at the Singularity Summit 2007
- [Video URL, Transcript URL]

Excerpts:

- "Douglas Hofstadter said that artificial general intelligence is at least 100 years off, and Ben Goertzel says it's less than ten years."
- "Ray Kurzweil [says] that progress is increasing exponentially [...] but [Jonathan Huebner says] it peaked around 1900."
- "there is a 0.26 correlation between overconfidence and number of Google hits on your name" [regarding prediction of the future]