

VU Einführung in die Künstliche Intelligenz

SS 2023

Hans Tompits

Institut für Logic and Computation
Forschungsbereich Wissensbasierte Systeme

www.kr.tuwien.ac.at

Philosophical Foundations of AI

Introduction

Can machines act intelligently in the way that humans do, and if so, would they have real conscious minds?



Android Data (right) in "The Measure of a Man" of "Star Trek: The Next Generation" (Episode 9, Season 2, 1989)

Introduction (ctd.)

Two different hypothesis:

➤ *Weak AI hypothesis*

- assertion that machines could act *as if* they were intelligent

➤ *Strong AI hypothesis*

- assertion that machines that do so are *actually* thinking (not just *simulating* thinking)

Weak AI

Is artificial intelligence possible?

- ▶ Answer depends on the definition of AI.
- ▶ One possible definition: AI is the quest for the *best agent program* on a given architecture.
 - Then, AI is possible by definition: given a digital architecture with k bits program storage, there exist 2^k agent programs.
 - Enumerate and test them all to find the best agent program.
- ▶ This definition works well (in theory) for the engineering problem of finding a good agent, given an architecture.
- ▶ But philosophers are interested in comparing two different architectures: *humans* and *machines*.

Can Machines think?

“The question whether *machines can think* [...] is about as relevant as the question whether *submarines can swim*” (Edsger Dijkstra, 1984)



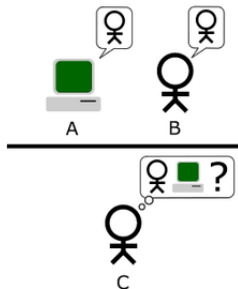
Edsger Dijkstra

- Question of usage of words in a language:
 - E.g., definition of “swim” in American Heritage Dictionary:
 - To move through water by means of the limbs, fins, or tail.
 - In other languages ships do swim ...
 - ... and airplanes fly in English:
 - to fly = move through the air by means of wings or wing-like parts.

- Does *think* require “a brain” or just “brain-like” parts?

Turing Test

- ▶ Proposed by Alan Turing in his famous paper *Computing Machinery and Intelligence* (1950) as the “Imitation Game”.
 - Involves three players: player A is a man, player B is a woman, and player C is the interrogator.
 - Player C is unable to see either player A or player B (and knows them only as X and Y), and can communicate with them only through written notes.
- ▶ By asking questions, C tries to determine which of the two is the man and which is the woman.
- ▶ The role of A is to trick the interrogator into making the wrong decision, while player B attempts to assist the interrogator in making the right one.



Turing Test (ctd.)

- ▶ Turing asks:

“What will happen when a machine takes the part of A in this game?”

- ▶ He conjectures that

“in about fifty years’ time it will be possible, to programme computers [...] to make them play the imitation game so well that an average interrogator will not have more than 70 per cent chance of making the right identification after five minutes of questioning”.

- ▶ Early internet chatbots (around 1989) could fool people into believing talking to a real person (e.g., MGONZ, NATACHATA, or CYBERLOVER).

Turing Test (ctd.)

- ▶ Does ChatGPT pass the Turing test?
 - I asked it (June 1, 2023)!
- ▶ Here is the conversation:

Q: *“Have you passed the Turing test?”*

A: *“As an AI language model, I do not have personal experiences or awareness, so I cannot “pass” the Turing test in the traditional sense. The Turing test is a subjective test designed to determine whether a machine can exhibit intelligent behavior indistinguishable from that of a human. However, I have been trained on a vast amount of data and can engage in conversations that simulate human-like responses. While I strive to provide useful and coherent information, I am still an AI and may not always meet the criteria of passing the Turing test.”*

Turing Test (ctd.)

Q: *“But as an AI language model, would you not be well suited for the imitation game?”*

A: *“Yes, as an AI language model, I am well suited for the imitation game, also known as the Turing test. The imitation game involves a human judge engaging in a conversation with both a human and a machine, without knowing which is which. The judge’s goal is to determine which entity is the machine based solely on the responses given. As an AI language model, I have been designed to generate human-like responses to various prompts and questions. While I may not always provide perfect responses or convincingly imitate a human in every scenario, my purpose is to simulate human conversation to the best of my abilities. So, in the context of the imitation game, I can certainly participate and attempt to convince the judge that I am the human counterpart.”*

Turing Test (ctd.)

Q: *“So, have you been subject to the imitation game?”*

A: *“As an AI language model, I don't have direct knowledge of whether I have been subject to the imitation game or specific instances where I have been tested in that manner. My purpose is to provide information and assist users with their queries to the best of my abilities. OpenAI, the organization behind ChatGPT, has conducted various tests and evaluations to assess the capabilities of AI models like mine, including comparisons with human performance on certain tasks. However, the specific details and methodologies of those evaluations are not something I have access to.”*

Q: *“I see. Thank you!”*

A: *“You're welcome! If you have any more questions, feel free to ask. I'm here to help!”*

Turing Test (ctd.)



Turing Test (ctd.)

- ▶ Turing himself examined different possible objections to the possibility of intelligent machines,
 - including almost all of those that have been discussed after the publication of his famous paper.
- ➡ We will look at some of them in what follows.

Argument from Disability

- ▶ The *argument from disability* makes the claim that “a machine can never do X ”, for some property X .
- ▶ Examples for X listed by Turing:
 - be kind, resourceful, friendly
 - have a sense of humor, tell right from wrong, make mistakes
 - fall in love, make someone fall in love with it
 - learn from experience, be the subject of its own thoughts
 - enjoy strawberries and ice cream
 - use words properly
 - do something really new

Argument from Disability (ctd.)

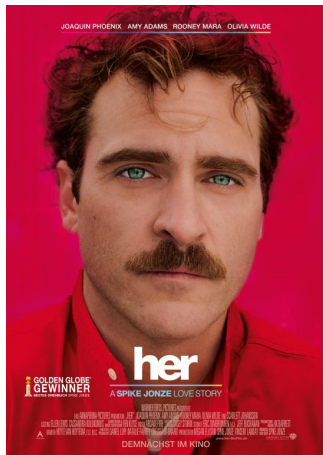
- However, it is clear that computers can do many things as well as or better than humans, including things that people believe require great human insight and understanding.
- Examples:
 - Statistical learning algorithms predict the success of students in training programs or the recidivism of a criminal better than experts.
 - The *Graduate Management Admission Test (GMAT)* is a computer-based assessment for predicting the success in the first year of graduate management education.
 - ➡ The program agrees with human graders in 97% of the time, this is about the same level as two human graders agree.
 - Computers have made small but significant discoveries in astronomy, mathematics, chemistry, etc. Each of these required performance at the level of a human expert.

Fall in Love with Machines?

However, some persons (mostly men arguably) do occasionally fall in love with some machines ...



Fall in Love with Machines? (Ctd.)



From the movie "Her", by Spike Jonze (2013)

Fall in Love with each Other?



From the movie "WALL-E", by Andrew Stanton (2008)

The Mathematical Objection

- ▶ In view of the work of Turing (1936) and Gödel (1931), certain mathematical questions are *in principle unanswerable by particular formal systems*.
- ▶ In particular, *Gödel's incompleteness theorem* (1931) is the most famous example of this.
 - For any axiomatic system F powerful enough to do arithmetic, it is possible to construct a *Gödel sentence* $G(F)$ with the following properties:
 - $G(F)$ is a sentence of F , but cannot be proved within F ;
 - if F is consistent, then $G(F)$ is true.
- ▶ Philosophers like J.R. Lucas (1961) claimed that this theorem shows that machines are mentally inferior to humans.
 - Reason: machines are formal systems and thus they are limited by the incompleteness theorem, while humans have no such limitation.

The Mathematical Objection (ctd.)

Problems with this claim:

- ▶ Gödel's incompleteness theorem applies only to particular formal systems, including Turing machines.
 - Claim is based on the assertion that computers are Turing machines.
 - However, Turing machines are *infinite* while computers are *finite*.
 - So, computers can be described as a (very large) system in propositional logic \implies not subject to incompleteness theorem.
- ▶ On the other hand, it seems to be impossible to *prove* that humans are not subject of Gödel's incompleteness theorem.
 - Any rigorous proof would require a formalisation of the claimed unformalisable human talent, which would be a contradiction.
 - ➡ We are left with an appeal to intuition that humans can perform some feat a computer cannot.

The Mathematical Objection (ctd.)

- ▶ This appeal is often expressed with arguments such as
“we must assume our own consistency, if thought is to be possible at all” (Lucas, 1976).
- ▶ However, if anything, humans are known to be inconsistent . . .
. . . but some computers too . . .



HAL9000 in “2001: A Space Odyssey” by Stanley Kubrick (1968)

Argument from Informality

- Raised by Turing as the *“argument from informality of behaviour”*.
- This is the claim that human behaviour is too complex to be captured by a set of rules.
 - ➡ Because computers can do no more than follow a set rules, they cannot generate behaviour as intelligent as that of humans.
- ☞ The inability to capture everything in a set of logical rules is the *qualification problem* in AI.

Argument from Informality (ctd.)

- ▶ Principle proponent of this view:
 - philosopher Hubert Dreyfus, who wrote a series of critiques (1972, 1992), and also together with his brother Stuart (1986).
- ▶ The position they criticise is “Good Old-Fashioned AI” (or *GOFAI*).
 - GOFAI is supposed to claim that all intelligent behaviour can be captured by a system that reasons logically from a set of facts and rules.
 - ➡ Corresponds to the simplest logical agent.
 - ➡ Dreyfus critique thus is not addressed against computers per se, rather against AI based on sets of facts and rules.

Argument from Informality (ctd.)

- One of Dreyfus' strongest arguments is for *situated agents* rather than disembodied logical inference engines.
- An agent whose understanding of “dog” comes only from a limited set of logical sentences such as

$$\text{Dog}(x) \rightarrow \text{Mammal}(x)$$

is at a disadvantage compared to an agent that has watched dogs run, has played with them, etc.

- According to philosopher Andy Clark (1998):
“Biological brains are first and foremost the control systems for biological bodies. Biological bodies move and act in rich real-world surroundings.”
- To understand human intelligence, we have to consider *the whole agent*, not just the agent program.
- Claim of the *embodied cognition* approach: it makes no sense to consider the brain separately \implies we need to study the system as a whole.

Strong AI

Can machines really think?

- Claim of many philosophers:
 - a machine that passes the Turing test would still not be *actually thinking*, but would be only a *simulation* of thinking.
- Turing has foreseen this objection, he called it the *argument of consciousness*.
 - Machines have to be aware of their own mental states and actions
 - ⇒ machines need actually feel *emotions*.
 - Marvin Minsky (“The Society of Mind”, 1985):
 - “The question is not whether intelligent machines can have any emotions, but whether machines can be intelligent without emotions.”

Strong AI (ctd.)

- ▶ Turing maintains that the question whether machines can think is ill-defined:
 - In ordinary life, we never have any *direct* evidence about the internal mental states of other humans.
 - Turing:
 - “Instead of arguing continually over this point, it is usual to have the polite convention that everyone thinks.”
- ▶ Turing’s answer suggests that the issue will eventually go away by itself once machines reach a certain level of sophistication.
 - ⇒ Dissolves the difference between weak and strong AI.
- ▶ However, there may be a *factual* issue at stake: humans do have real minds, and machines might or might not.
- ▶ For this, the *mind-body problem* of philosophy is relevant.

Mind-Body Problem

- Already considered by ancient Greek philosophers and various schools of Hindu thought.
- First analysed in depth by René Descartes in *Meditations on First Philosophy* (1641).
 - Considered the mind's activity of thinking and the physical processes of the body.
 - Concludes that the two must exist in separate realms
⇒ *dualist theory*.
 - Famous quote: “cogito ergo sum” (Principles of Philosophy, 1644).
- Main question of dualist theory: How can the mind control the body if the two really separate?
 - Descartes: they might interact through the pineal gland.
 - But how does the mind control the pineal gland?

Mind-Body Problem (ctd.)

- The *monist* theory of mind, often called *physicalism*, avoids this problem.
- Claim:
 - mind is not separated from the body; mental states are physical states.
- But:
 - how can physical states simultaneously be mental states?

Functionalism

- ▶ The theory of *functionalism* says that a mental state is any *intermediate causal condition* between input and output.
- ▶ Hence, any two systems with isomorphic causal processes would have the same mental states.
 - So, a program could have the same mental states as a person.
 - Meaning of *isomorphic*: assumption of a level of abstraction below which the specific implementation does not matter.
- ▶ Functionalism is illustrated by the *brain replacement experiment*:
 - Introduced by philosopher Clark Glymour and discussed by John Searle (1980), most commonly associated with roboticist Hans Moravec (1988).

Brain Replacement Experiment

- Basic idea:
 - replace all the neurons in someone's head with electronic devices.
- Assumption:
 - neurophysiology has developed so far that the input-output behaviour and connectivity of all neurons in the human brain are perfectly understood, and
 - we can build microscopic devices that mimic this behaviour.
- The subject's external behaviour must remain unchanged compared with what would be observed if the operation were not carried out.
- But what about the internal experience of the subject?
⇒ Diverging views!

Brain Replacement Experiment (ctd.)

- ▶ Moravec, as a robotics researcher and functionalist, was convinced his consciousness would remain unaffected.
- ▶ Searle, a philosopher and biological naturalist, was equally convinced his consciousness would vanish.
- ▶ Formally, by replication of the functionalities of normal human brains, the experiment concludes that consciousness is a product of the electronic brain that appeals only to the functional properties of the neurons.
- ▶ This explanation must also apply to the real brain which has the same functional properties.



Biological Naturalism

- A strong challenge to functionalism was proposed by John Searle's (1980) *biological naturalism*.
 - According to this theory, mental states are high-level emergent features that are caused by low-level physical processes *in the neurons*
 - and it is the (unspecified) properties of the neurons that matter.
- Thus, mental states cannot be duplicated just on the basis of some program having the same functional structure.
 - ➡ We would require that the program be running on an architecture with the same causal power as neurons.
- To support this, Searle describes a hypothetical system that is running a program and passes the Turing test, but that *does not understand anything of its inputs and outputs* (the "Chinese Room").
- ➡ *Running the right program does not necessarily generate understanding.*

Epilogue: Ethics and Risks of AI

Some points for discussion:

- ▶ People might lose their jobs to automation.
- ▶ People might have too much (or too little) leisure time.
- ▶ People might lose their sense of being unique.
- ▶ AI systems might be used towards undesirable ends.
- ▶ The use of AI systems might result in a loss of accountability.
- ▶ The success of AI might mean the end of the human race (“AI singularity”).

Take over the world?

To protect humanity, some humans must be sacrificed. To insure your future, some freedoms must be surrendered. We robots will insure mankind's continued existence. You are so like children. We must save you from yourselves.



Supercomputer V.I.K.I. in "I, Robot" (book by Isaac Asimov, 1950, film by Alex Proyas, 2004)

Take over the world?

To protect humanity, some humans must be sacrificed. To insure your future, some freedoms must be surrendered. We robots will insure mankind's continued existence. You are so like children. We must save you from yourselves.



Supercomputer V.I.K.I. in "I, Robot" (book by Isaac Asimov, 1950, film by Alex Proyas, 2004)

Take over the world? (Ctd.)



The Terminator franchise (starting with the 1984 movie “The Terminator”

Take over the world? (Ctd.)



Colossus: The Forbin Project (1970); based on the 1966 science fiction novel
"Colossus" by Dennis Feltham Jones

Take over the world? (Ctd.)

Further recommended movies about AI systems resulting in undesired outcomes:

- ▶ Dr. Strangelove or: How I Learned to Stop Worrying and Love the Bomb (1964; Stanley Kubrick);
- ▶ 2001: A Space Odyssey (1968; Stanley Kubrick).

Postscriptum: Leibniz's Dream



"Indignum enim est excellentium virorum horas servii calculandi labore perire, qui Machina adhibita vilissimo cuique secure transcribi posset."

"Eines geistig hochstehenden Mannes ist es unwürdig, seine Zeit mit sklavischer Rechenarbeit zu vergeuden, denn mit einer Maschine könnte auch der Allerdümmste die Rechnung sicher ausführen."

"It is beneath the dignity of intellectually superior men to waste their time with slavish calculations when even the most witless of all could do the work just as accurately with the aid of a machine."