Introduction to Artificial Intelligence

Roman Barták
Department of Theoretical Computer Science and Mathematical Logic

Four approaches to AI

We approached artificial intelligence as the science of designing rational agents – agents that maximize own expected utility.

There are other views of AI:

	human performance	*ideal performance*
though processes	**Thinking humanly:** The cognitive modeling approach (psychology, neuroscience)	**Thinking rationally:** The "laws of thought" approach (logic)
behavior	**Acting humanly:** The Turing test approach (behaviorism)	**Acting rationally:** The rational agent approach

Empirical science with observations and hypothesis about human behavior

Combination of **mathematics and engineering** (using computer as formal reasoning tools)

Human utility

Universal exchangeability of **money** for all kinds of goods and services suggests that money plays a significant role in human utility functions.

But this does not mean that money behaves as a **utility function**.

Assume that you won a competition and the host offers you a choice: either you can take the 1 mil. USD price or you can gamble it on the flip of coin. If the coin comes up heads, you end up with nothing, but if it comes up tails, you get 2.5 mil. USD. What is your choice?

- – Expected monetary value of the gamble is 1 250 000 USD.
- – Most people decline the gamble and pocket the million. Is it irrational?

The decision does not depend on the prize only but also on the wealth of the player!

Let S_n denote a state of possessing total wealth **n** USD, and the current wealth is **k** USD.

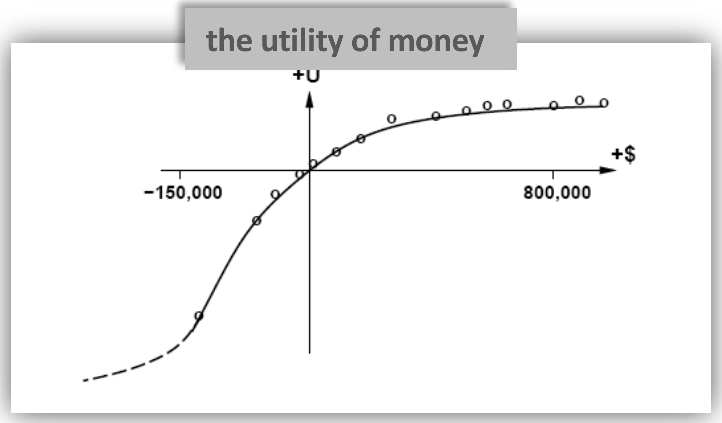
Then the expected utilities of two actions are:

- EU(Accept) = ½ $U(S_k)$ + ½ $U(S_{k+2.500.000})$
- EU(Decline) = $U(S_{k+1.000.000})$

Suppose we assign:

- $U(S_k) = 5$, $U(S_{k+1.000.000}) = 8$, $U(S_{k+2.500.000}) = 9$.

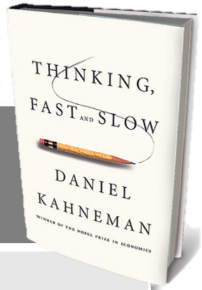
Then the decision to take the million is rational.



the utility of money

Human judgment

The evidence suggests that humans are "predictably irrational ".



Allais paradox

What is your choice?

A: 80% chance of winning 4000 USD B: 100% chance of winning 3000 USD	C: 20% chance of winning 4000 USD D: 25% chance of winning 3000 USD
Most people consistently prefer B over A (taking the sure thing!)	Most people prefer C over D (higher expected monetary value)

Certainty effect – people are strongly attracted to gains that are certain.

Ellsberg paradox

The urn contains 1/3 red balls, and 2/3 either black or yellow balls. What is your choice?

A: win 100 USD for a red ball B: win 100 USD for a black ball	C: win 100 USD for a red or yellow ball D: win 100 USD for a black or yellow ball
Most people prefer A over B (A gives a 1/3 chance of winning, while B could be anywhere between 0 and 2/3)	Most people prefer D over C (D gives you a 2/3 chance, while C could anywhere between 1/3 and 3/3)

If you think there are more red than black balls, then you should prefer A over B and C over D.

Ambiguity aversion – most people elect the known probability rather than the unknown unknown.

Weak AI: machines can act as if they were intelligent

Most AI researchers take weak AI hypothesis as granted.

Strong AI: machines are actually thinking (not just simulating thinking)

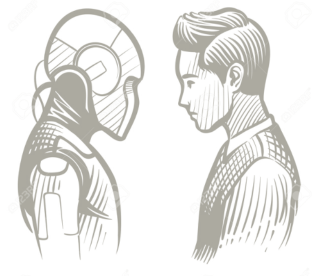
It is not clear if it is possible at all (but is it important?).

General AI: machines can solve any problem (like a human)

Strong AI is in principle general AI, but can weak AI be also general AI?

Narrow AI: machines solve problems constrained to a specific area

Granted for sure.



Turing test

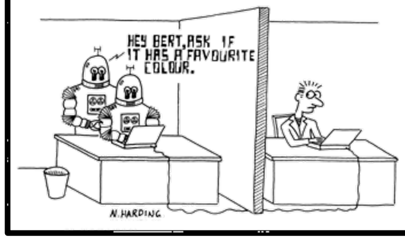
Instead of asking whether machines can think, we should ask whether machines can pass a behavioral test.

- the computer program has a conversation with an interrogator (for five minutes)
- the interrogator has to guess if the conversation is with a program or a person (and the computer fools the interrogator in 30% of cases)

Has a computer already passed the Turing test?

Many people have been fooled when they didn’t know they might be chatting with a computer:

- *ELIZA* program, computer “psychotherapist” (1966, Joseph Weizenbaum)
- *Eugene Goostman*, chatbot (2012, winner of The Loebner Prize)
- *Google Duplex*, natural conversations to carry out “real world” tasks over the phone (2018)



Reverse Turing test

computer attempts to recognize whether it communicates with a computer or a person

Gödel's incompleteness theorem

For any formal axiomatic system F powerful enough to do arithmetic, it is possible to construct a so-called 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.

Is Gödel's incompleteness theorem a problem for AI?

- Gödel's incompleteness theorem applies to formal systems that are powerful enough to do arithmetic (this includes Turing machines). But is computer a Turing machine? Computers are finite and they can be described as a (very large) system in proposition logic (which is not subject to Gödel's incompleteness theorem).
- An agent should not be too ashamed that it cannot establish the truth of some sentence while other agents can. ("I cannot consistently assert that this sentence is true").
- Even if computers have limitations of what they can prove, there is no evidence than humans are immune from those limitations (humans are known to be inconsistent).

Dualist theory
mind and body exist in separate realms

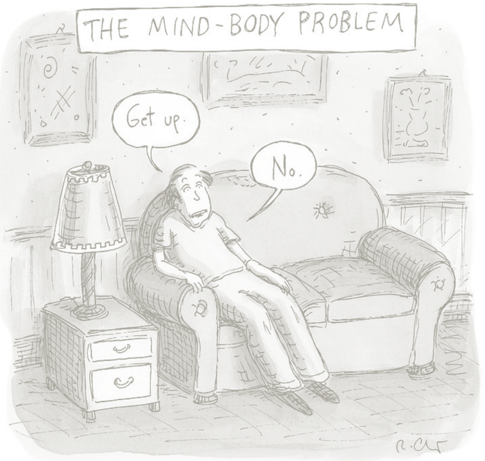
– How can the mind control the body?

Monist theory (physicalism)
mind is not separate from body

– Mental states are physical states.

– Prevailing approach today.

– Possibility of strong AI.



Though experiment: Brain in a vat

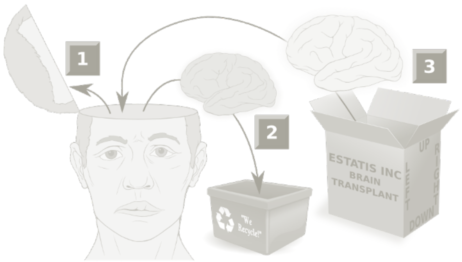
Imagine that brain is removed from body at birth and placed in a marvelously engineered vat. Electronic signals are fed to the brain from a computer simulation of an entirely fictitious world and motor signals from the brain are used to modify the simulation.

Is eating a hamburger in simulation different from eating real hamburger?

Though experiment: Brain replacement

Suppose that we perfectly understand input-output behavior and connectivity of all the neurons in the human brain and we can replace individual neurons with corresponding electronic devices without interrupting the operation of the brain as a whole.

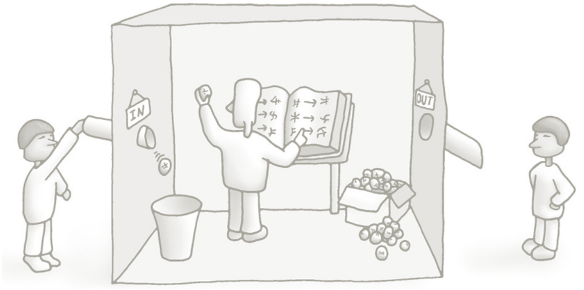
Will the subject gradually lose consciousness during above operation while the external behavior remains the same?



Though experiment: Chinese room

Assume that you sit in a room with a book of instructions, you receive some symbols on one side, follow instructions in the book based on symbols received, and eventually draw some symbols that you pass back to the outside world. From outside, we see a system that is taking input in the form of Chinese sentences and generating answers in Chinese.

Do you understand Chinese?



Should we develop AI?

- people may lose trust (fake information)
- people may lose their privacy (surveillance)
- people may lose their jobs (automation)
- people may lose their lifes (killing machines)
- the end of human race (singularity)

Explainability

What if the human wants to know why the AI system generated a given output?

Fragility

What if we slightly change the input to the AI system and the AI system generates a dramatically different output?

Bias

What if the AI system gives advantage to some group of users?

AI: social issues

Military usage

Can AI system trigger the fire button?

Employment

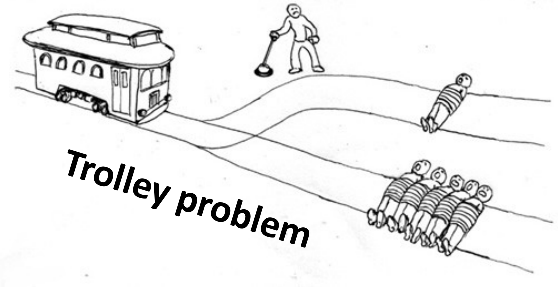
Will AI systems automate every human task?

Surveillance

Can AI systems indirectly control human lives?

Decision making

Should AI systems do the decisions
(or recommend only)?



Trolley problem

Department of Theoretical Computer Science and Mathematical Logic
bartak@ktiml.mff.cuni.cz