

26. Philosophical Foundations

Strong AI vs Weak AI

- Philosophers' Concerns
 - Is it possible for machines to act intelligently in the way that people do, and if they did, would they have real, conscious minds?
- Weak AI Hypothesis
 - Machines can **act as if they were intelligent**
- Strong AI Hypothesis
 - Machines that act intelligent **are actually thinking** (not just simulating thinking)
- What do AI researchers think?
 - Take the weak AI hypothesis for granted
 - Don't care about the strong AI hypothesis
 - as long as their program works, they don't care whether you call it a simulation of intelligence or real intelligence



26.1 Weak AI: Can Machines Act Intelligently?

AI was founded on the following assumption (i.e. assuming that weak AI is possible):

“Every aspect of learning or any other feature of intelligence can be so precisely described that a machine can be made to simulate it.” - 1955

- Engineers
 - The definition of AI works well for the engineering problem of finding a good agent, given an architecture
- Philosophers
 - They are interested in the problem of comparing two architectures—human and machine
 - They pose the “weak AI question” not in terms of maximizing expected utility but rather as, “Can machines think?”

26.1 Can Machines Think?

“The question of whether Machines Can Think . . . is about as relevant as the question of whether Submarines Can Swim.” - Edsger Dijkstra (1984)

Example 1:

- Swim = To move through water by means of the limbs, fins, or tail (Dictionary Meaning)
- Most people agree that submarines, being limbless, cannot swim

Example 2:

- Fly = To move through the air by means of wings or winglike parts (Dictionary Meaning)
- Most people agree that airplanes, having winglike parts, can fly

These two (opposite) examples are not concerned with design or capabilities of airplanes and submarines, rather they are about the usage of words in English

26.1 Can Machines Think?

What is “think”? - does it require “a brain” or just “brain-like parts”

think

/THiNGk/ 🔊

verb

verb: **think**; 3rd person present: **thinks**; past tense: **thought**; past participle: **thought**; gerund or present participle: **thinking**

1. have a particular opinion, belief, or idea about someone or something.
"she thought that nothing would be the same again"
synonyms: **believe**, be of the opinion, be of the view, be under the impression; [More](#)
 - used in questions to express anger or surprise.
"What do you think you're doing?"
 - used in speech to reduce the force of a statement or opinion, or to politely suggest or refuse something.
"I thought we could go out for a meal"
2. direct one's mind toward someone or something; use one's mind actively to form connected ideas.
"he was thinking about Colin"
synonyms: **ponder**, **reflect**, **deliberate**, **consider**, **meditate**, **contemplate**, **muse**, **ruminate**, be lost in thought, be in a brown study, **brood**; [More](#)
 - take into account or consideration when deciding on a possible action.
"you can live how you like, but there's the children to think about"
 - consider the possibility or advantages of (a course of action).
"he was thinking of becoming a zoologist"
 - have a particular mental attitude or approach.
"he thought like a general"
 - have a particular opinion of.
"I think of him as a friend"

swim

/swim/ 🔊

verb

1. propel the body through water by using the limbs, or (in the case of a fish or other aquatic animal) by using fins, tail, or other bodily movement.

fly¹

/fli/ 🔊

verb

verb: **fly**; 3rd person present: **flies**; past tense: **flew**; gerund or present participle: **flying**; past participle: **flown**

1. (of a bird or other winged creature) move through the air under control.
"close the door or the moths will fly in"
synonyms: travel through the air, wing its way, [wing](#), [glide](#), [soar](#), [wheel](#); [More](#)

Conclusion:

The question “Can machines think?” is ill defined.

26.2 Strong AI: Can Machines Really Think?

- Philosophers:
 - Many philosophers have claimed that a machine that passes the Turing Test would still not be actually thinking, but would be only a simulation of thinking
 - In other words, **can a machine be conscious?** — can a machine be aware of its own mental states and actions?
- Turing's response:
 - Turing does not claim that machine CAN (in fact) be conscious
 - “The question is just as ill-defined as asking, “Can machines think?”
 - “Why should we insist on a higher standard for machines than we do for humans?”
 - In ordinary life we never have any direct evidence about the internal mental states of other humans
 - “Instead of arguing continually over this point, it is usual to have the polite convention that everyone thinks.”
- One can easily imagine some future time in which conversations with machines are commonplace, and it becomes customary to make no linguistic distinction between “real” and “artificial” thinking. Do you agree?



26.2 Strong AI: Can Machines Really Think?

In the context of artificial vs real thinking...

- History of Artificial Urea

- Organic and Inorganic chemistry were essentially disjoint enterprises
 - and many thought that no process could exist that would convert inorganic chemicals into organic material
- Artificial urea was synthesized for the first time by Frederick Wöhler in 1848
- Once the synthesis was accomplished, chemists agreed that artificial urea was urea, **because it had all the right physical properties**
- Those who had argued “an intrinsic property possessed by organic material cannot be possessed by inorganic material” were faced with the impossibility of devising any test that could reveal the supposed deficiency of artificial urea
- For thinking, AI has not yet reached the “1848” and there are those who believe that artificial thinking, no matter how impressive, will never be real



26.2 Strong AI: Can Machines Really Think?

Philosopher John Searle:

“No one supposes that a computer simulation of a storm will leave us all wet . . . Why on earth would anyone in his right mind suppose a computer simulation of mental processes actually had mental processes?”

Authors:

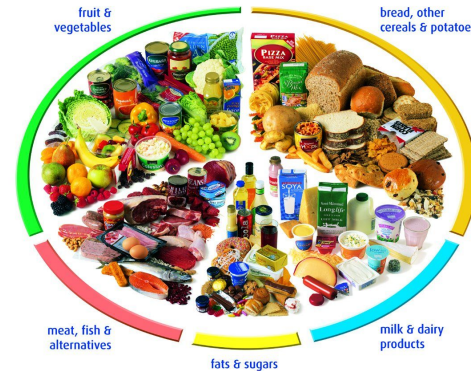
- While it is easy to agree that **computer simulations of storms** do not make us wet, but it is not clear how to carry this analogy over to **computer simulations of mental processes**
- A Hollywood simulation of a storm using sprinklers and wind machines does make the actors wet
- Most people are comfortable saying that a **computer simulation of addition is addition**, and of chess is chess
- We typically speak of **an implementation of addition or chess, not a simulation**

Are mental processes more like storms, or more like addition?



26.3 The Ethics and Risks of Developing AI

- We have largely concentrated on whether we can develop AI
 - but we must also consider whether we should
 - If the effects of AI technology are more likely to be negative than positive, then it would be the moral responsibility of workers in the field to redirect their research



For example, Should we develop applications that “better” advertise cigarettes or a healthy diet?

26.3 Will People Lose Their Jobs to Automation?



- Example 1
 - Much of the economy in the US depends on the availability of consumer credit
 - AI programs (currently) handle the credit card applications, charge approvals, and fraud detection
 - **Argument:** Thousands of workers have been displaced by these AI programs
 - **Counter-argument:** If we took the AI away these jobs would not exist because human labor would add an unacceptable cost to the transactions
- Example 2
 - In 2005, at [AAAI](#), Nils Nilsson set as a challenge the creation of **human-level AI that could pass the employment test** rather than the Turing test - a robot that could learn to do any one of a range of jobs
 - Examples, self-driving cars, skin cancer detection, student admission
- We may end up in a future where unemployment is high, but even the unemployed serve as managers of their own cadre of robot workers

26.3 Will People Lose Their Sense of Being Unique?

- AI research makes possible the idea that humans are (conscious) automata
 - an idea that results in a loss of autonomy or even of humanity
- But humanity has already survived (at least) two of such setbacks to our sense of uniqueness
 - (a) Copernicus (1543) moved the Earth away from the center of the solar system
 - (b) Darwin (1871) put Homo Sapiens at the same level as other species
- AI, if widely used, may be at least as threatening to the moral assumptions of 21st-century society as Darwin's theory of evolution was to those of 19th century

26.5 Will AI Be Used Towards Undesirable Ends?

“A science is said to be useful if its development tends to accentuate the existing inequalities in the distribution of wealth, or more directly promotes the destruction of human life.” - G. H. Hardy 1940

- But, this holds for all sciences, AI being no exception
 - Autonomous AI systems are now commonplace on the battlefield; the U.S. military deployed over 5,000 autonomous aircraft and 12,000 autonomous ground vehicles in Iraq [TED - P.W. Singer: Military robots and the future of war](#) 
- Military robots = Medieval armor [TED - Peter Haas: The Real Reason to be Afraid of Artificial Intelligence](#) 
 - Some argue that military robots are like medieval armors (taken to a extreme)
 - But, as human decision making is taken out of the firing loop, robots may end up making decisions that lead to the killing of innocent civilians
 - Possession of powerful robots (like the possession of sturdy helmets) may give a nation overconfidence, causing it to go on war more recklessly than necessary



≠

=



26.3 Will Use of AI Result in Loss of Accountability?

- Perspective 1

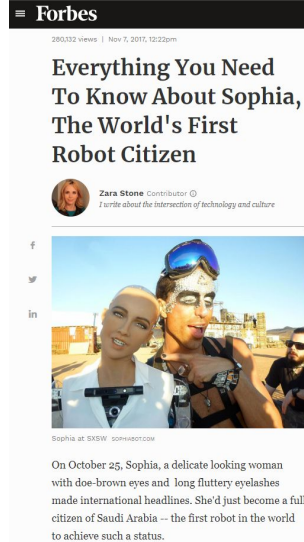
- When a physician relies on the judgment of a medical expert system for a diagnosis, who is at fault if the diagnosis is wrong?
 - If a physician performs medical procedures that have high expected utility, even if the actual result is catastrophic for the patient - it is now accepted that negligence cannot be shown
- “Who is at fault if a diagnosis is unreasonable?”
 - Courts have held that **medical expert systems play the same role as medical textbooks** and reference books
 - i.e. physicians are responsible for understanding the reasoning behind any decision and for using their own judgment in deciding whether to accept the system’s recommendations

- Perspective 2

- If AI systems become reliably more accurate than human diagnosticians, doctors might become legally liable if they **DON'T** use the recommendation of an AI system (already becoming true!!)

26.3 Will Use of AI Result in Loss of Accountability?

- Monetary transactions by an AI system
 - If monetary transactions are made “on one’s behalf” by an intelligent agent, is one liable for the debts incurred? Would it be possible for an intelligent agent to have assets itself and to perform electronic trades on its own behalf?
 - Book - “No program has been granted legal status as an individual for the purposes of financial transactions; at present, it seems unreasonable to do so” (already becoming true!!)
- Can we consider a program as a “driver”?
 - Programs will need to be considered “drivers” for the purposes of enforcing traffic regulations on real highways
 - In California law, at least, there do not seem to be any legal sanctions to prevent an automated vehicle from exceeding the speed limits, although the designer of the vehicle’s control mechanism would be liable in the case of an accident
 - As with human reproductive technology, the law has yet to catch up with the new developments (already becoming true!!)



26.3 Does Success of AI Mean End of Human Race?



Which of these do you think are the greatest threats to end human race? Rank them.

1. An AI system's state estimation may be incorrect, causing it to do the wrong thing. For example, an autonomous car might incorrectly estimate the position of a car in the adjacent lane, leading to an accident that might kill the occupants. More seriously, a missile defense system might erroneously detect an attack and launch a counterattack, leading to the death of billions. These risks are not really risks of AI systems—in both cases the same mistake could just as easily be made by a human as by a computer. The correct way to mitigate these risks is to design a system with checks and balances so that a single state-estimation error does not propagate through the system unchecked.
2. Specifying the right utility function for an AI system to maximize is not so easy. For example, we might propose a utility function designed to minimize human suffering, expressed as an additive reward function over time. Given the way humans are, however, we'll always find a way to suffer even in paradise; so the optimal decision for the AI system is to terminate the human race as soon as possible—no humans, no suffering.
With AI systems, then, we need to be very careful what we ask for, whereas humans would have no trouble realizing that the proposed utility function cannot be taken literally.
3. "Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an "intelligence explosion," and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control." - I. J. Good (1965)

The "intelligence explosion" has also been called the technological singularity by mathematics professor and science fiction author Vernor Vinge, who writes (1993), "Within thirty years, we will have the technological means to create superhuman intelligence. Shortly after, the human era will be ended."

So far, every other technology has followed an S-shaped curve, where the exponential growth eventually tapers off. Sometimes new technologies step in when the old ones plateau; sometimes we hit hard limits. With less than a century of high-technology history to go on, it is difficult to extrapolate hundreds of years ahead.

4. Science fiction writer Isaac Asimov (1942) was the first to address this issue, with his three laws of robotics: (1) A robot may not injure a human being or, through inaction, allow a human being to come to harm (2) A robot must obey orders given to it by human beings, except where such orders would conflict with the First Law (3) A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.
These laws seem reasonable, at least to us humans. But the trick is how to implement these laws. In the Asimov story Roundabout a robot is sent to fetch some selenium. Later the robot is found wandering in a circle around the selenium source. Every time it heads toward the source, it senses a danger, and the third law causes it to veer away. But every time it veers away, the danger recedes, and the power of the second law takes over, causing it to veer back towards the selenium. The set of points that define the balancing point between the two laws defines a circle. This suggests that the laws are not logical absolutes, but rather are weighed against each other, with a higher weighting for the earlier laws.
5. Can we design robots that have static utility functions (such as, "never hurt babies and a person on a wheelchair")? We can't just give a program a static utility function, because circumstances, and our desired responses to circumstances, change over time. For example, if technology had allowed us to design a super-powerful AI agent in 1800 and endow it with the prevailing morals of the time, it would be fighting today to reestablish slavery and abolish women's right to vote.

Summary

- Philosophers use the term weak AI for the hypothesis that machines could possibly behave intelligently, and strong AI for the hypothesis that such machines would count as having actual minds (as opposed to simulated minds)
- We identified six potential threats to society posed by AI and related technology; we concluded that some of the threats are either unlikely or differ little from threats posed by “unintelligent” technologies
- Ultraintelligent machines might lead to a future that is very different from today—we may not like it, and at that point we may not have a choice