Philosophy and Ethics of AI

Louis Vervoort, PhD

Higher School of Economics
September-October 2023

The foundations of AI – a short historical overview, from R&N

4. Neuroscience.

- How do brains process information?

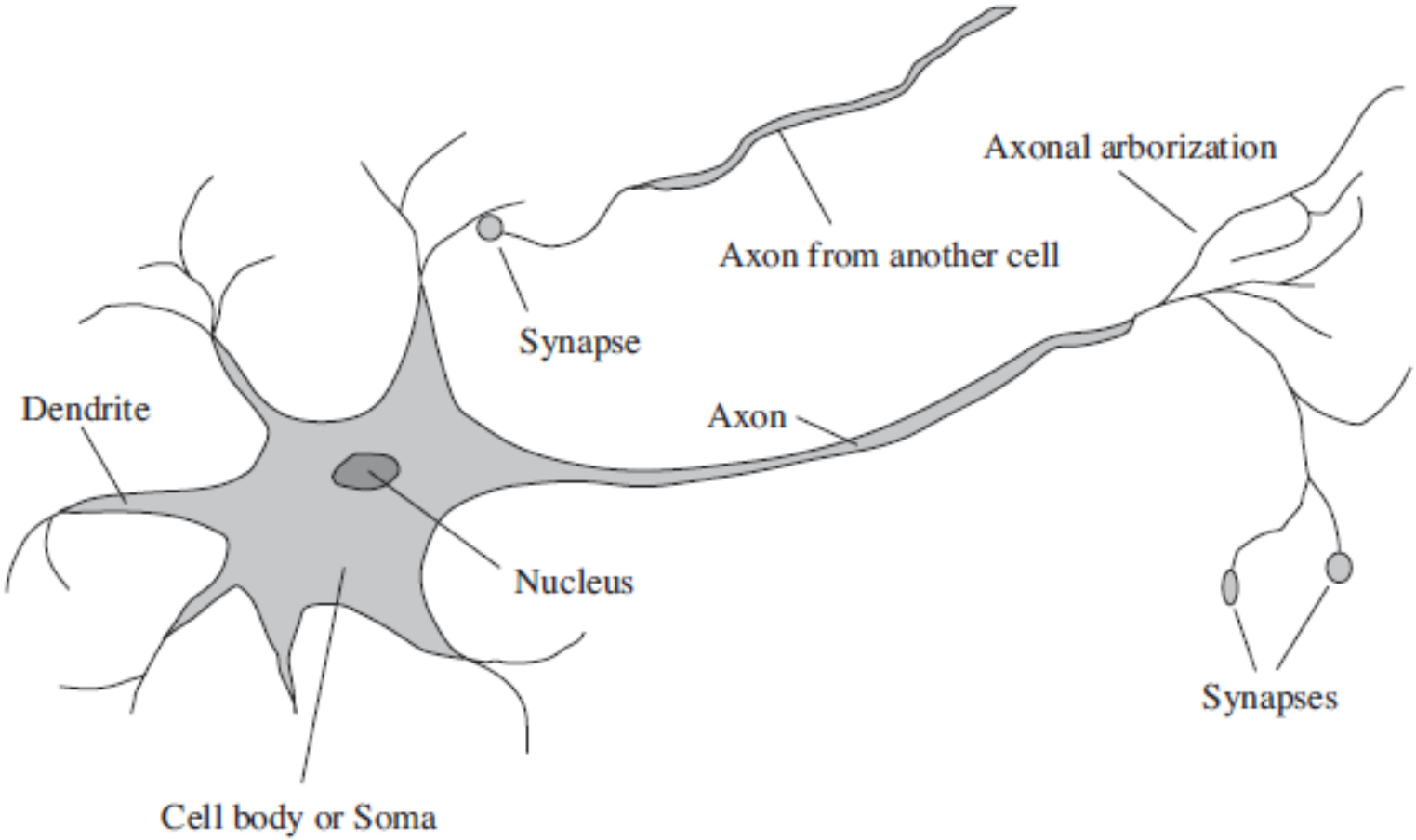
- Thinking is done by the brain (known since antiquity), but the exact way in which the brain enables thought is one of the great mysteries of science.
- Many people say the brain is the most complex system in the universe.
- In about 335 B.C. Aristotle wrote, "Of all the animals, man has the largest brain in proportion to his size."

Treeshrew,
record holder

The foundations of AI – a short historical overview, from R&N

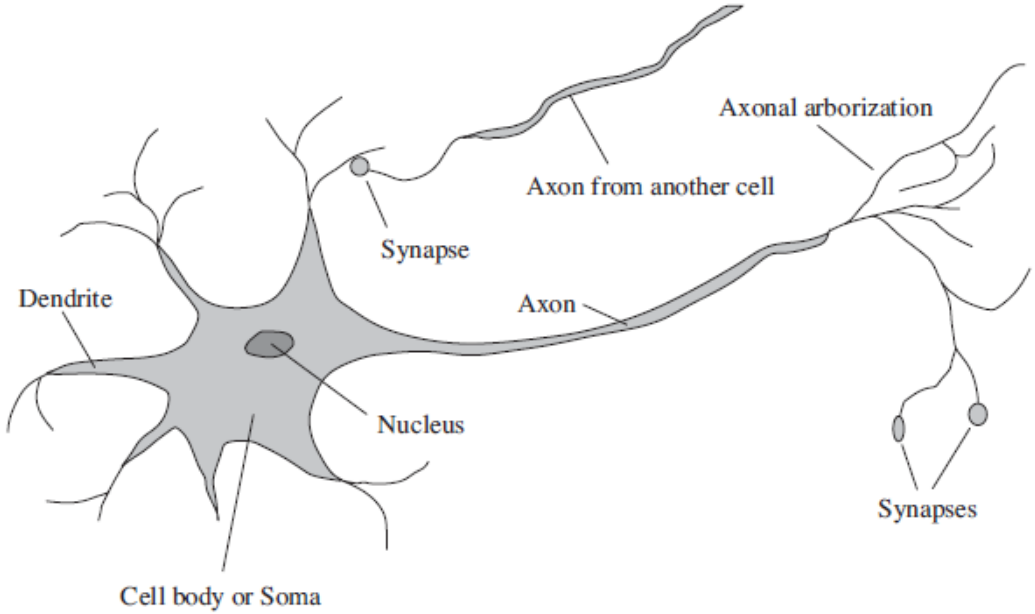
- Paul Broca's (1824–1880) study of aphasia (speech deficit) in brain-damaged patients in 1861 demonstrated the existence of localized areas of the brain responsible for specific cognitive functions.
- 1873: Camillo Golgi developed a staining technique allowing the observation of individual neurons.
- Santiago Ramon y Cajal (1852-1934): neuronal structures.
- Both Nobel prize in 1906.

The foundations of AI – a short historical overview, from R&N



Norvig, P. and Russell, S. (2010). Artificial Intelligence: A Modern Approach (3rd Edition).

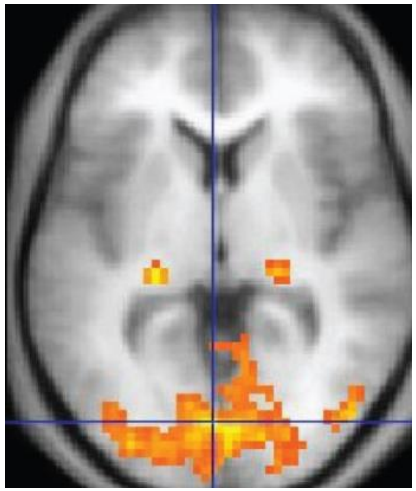
The foundations of AI – a short historical overview, from R&N



From R&N: The parts of a nerve cell or neuron. Each neuron consists of a cell body, or soma, that contains a cell nucleus. Branching out from the cell body are a number of fibers called dendrites and a single long fiber called the axon. The axon stretches out for a long distance, much longer than the scale in this diagram indicates. Typically, an axon is 1 cm long (100 times the diameter of the cell body), but can reach up to 1 meter. A neuron makes connections with 10 to 100,000 other neurons at junctions called synapses. Signals are propagated from neuron to neuron by a complicated electrochemical reaction. The signals control brain activity in the short term and also enable long-term changes in the connectivity of neurons. These mechanisms are thought to form the basis for learning in the brain. Most information processing goes on in the cerebral cortex, the outer layer of the brain. The basic organizational unit appears to be a column of tissue about 0.5 mm in diameter, containing about 20,000 neurons and extending the full depth of the cortex about 4 mm in humans).

The foundations of AI – a short historical overview, from R&N

- We now have some data on the mapping between areas of the brain and the parts of the body that they control or from which they receive sensory input.
- Brain is highly 'plastic' and flexible (interchangeable connections), but no solid theories of mechanisms.
- Almost no theory on how an individual memory is stored.



- Experimental techniques:
  - electroencephalography (EEG), Hans Berger, 1930
  - functional magnetic resonance imaging (fMRI), 1990ies
- R&N: "The truly amazing conclusion is that a collection of simple cells can lead to thought, action, and consciousness or, in the words of John Searle (1992), 'brains cause minds'."
- Alternative: mysticism: mind beyond science.

❖ **Brain is highly 'plastic' and flexible (interchangeable connections), but no solid theories of mechanisms.**

Neuroplasticity

30 languages

Article Talk | Read Edit View history Tools

From Wikipedia, the free encyclopedia

"Neural plasticity" redirects here. For journal, see Neural Plasticity (journal). For the 2014 Cold Specks album, see Neuroplasticity (album).

Neuroplasticity, also known as **neural plasticity**, or **brain plasticity**, is the ability of neural networks in the brain to change through growth and reorganization. It is when the brain is rewired to function in some way that differs from how it previously functioned.[1] These changes range from individual neuron pathways making new connections, to systematic adjustments like cortical remapping or neural oscillation. Other forms of neuroplasticity include homologous area adaptation, cross modal reassignment, map expansion, and compensatory masquerade.[2] Examples of neuroplasticity include circuit and network changes that result from learning a new ability, information acquisition, environmental influences, practice, and psychological stress.[3][4][5][6][7][8]

Neuroplasticity was once thought by neuroscientists to manifest only during childhood,[9][10] but research in the latter half of the 20th century showed that many aspects of the brain can be altered (or are "plastic") even through adulthood.[11] However, the developing brain exhibits a higher degree of plasticity than the adult brain.[12] Activity-dependent plasticity can have significant implications for healthy development, learning, memory, and recovery from brain damage.[13][14][15]

❖ **Is this type of plasticity (or even a simple form of it) implemented / implementable in AI ??**

The foundations of AI – a short historical overview, from R&N

- **R&N: "Brains and digital computers have somewhat different properties."**

	Supercomputer	Personal Computer	Human Brain
Computational units	10^4 CPUs, 10^{12} transistors	4 CPUs, 10^9 transistors	10^{11} neurons
Storage units	10^{14} bits RAM	10^{11} bits RAM	10^{11} neurons
	10^{15} bits disk	10^{13} bits disk	10^{14} synapses
Cycle time	10^{-9} sec	10^{-9} sec	10^{-3} sec
Operations/sec	10^{15}	10^{10}	10^{17}
Memory updates/sec	10^{14}	10^{10}	10^{14}

Figure 1.3 A crude comparison of the raw computational resources available to the IBM BLUE GENE supercomputer, a typical personal computer of 2008, and the human brain. The brain's numbers are essentially fixed, whereas the supercomputer's numbers have been increasing by a factor of 10 every 5 years or so, allowing it to achieve rough parity with the brain. The personal computer lags behind on all metrics except cycle time.

- **Is this really a good comparison ?**

- **This is a (very) partial comparison, since the neural (and electro-chemical) structure of the brain is much more complex than that of a computer and an ANN.**

The foundations of AI – a short historical overview, from R&N

- **Is this really a good comparison ?**
- **This is a (very) partial comparison, since the neural (and electro-chemical) structure of the brain is much more complex than that of a computer and an ANN.**
- **R&N: "Even with a computer of virtually unlimited capacity, we still would not know how to achieve the brain's level of intelligence."**

Philosophy and Ethics of AI

Louis Vervoort, PhD

Higher School of Economics
September-October 2023

Syllabus

- **Schedule Book chapter presentations (OralTest1) and Research presentations (OralTest2)**

	Lectures (Tuesdays)	Book chapter pres (Seminar, Thursday)	Research presentation (Seminar, Thursday)
Week 1, Tue 5 Sept	1, 2	- - -	- - -
Week 2, Tue 12 Sept	3	Team 1, Chapter 1	- - -
Week 3, Tue 19 Sept	4	NO SEMINAR	NO SEMINAR
Week 4, Tue 26 Sept	5	Team 2, Chapter 2 WED 27 ! Team 3, Chapter 3 THUR 28 !	Team 7 WED 27 sept ! Team 6 THUR 28 sept !
Week 5, Tue 3 Oct	6 + Test1	Team 4, Chapter 4	Team 5
Week 6, Tue 10 Oct	7	Team 5, Chapter 5	Team 4
Week 7, Tue 17 Oct	8	Team 6, Chapter 6 Team 7, Chapter 7	Team 3 Team 2 Extra seminar on Wedn.
Week 8, Tue 24 Oct	- - -	- - -	Team 1 (on Tue 24/10, instead of lecture)
Week 9, Tue 31 Oct	Exam1	- - -	- - -

Syllabus

- **Teams, Group 1**

Team 1	(Асташкин Артемий Андреевич) (absent)	
	Гнездилова Вера Андреевна	**RT 5**
	Дмитриченко Камила Дмитриевна	
	Досаев Савелий Юрьевич	
Team 2	Камбачеков Тимур Алимович	**RT 3**
	Кульжик Степан Михайлович	
	Куракина Анеля Романовна	
	Лапко Дарья Андреевна	
Team 3	Луговцев Тимур Андреевич	**RT 2**
	Малышева Мария Александровна	
	Манякин Павел Дмитриевич	
	Мирзак Александр Сергеевич	
	Нигматуллина Диана Дмитриевна	
Team 4	Ноздрина Марина Викторовна	**RT 11**
	Петров Илья Родионович	
	Раташнюк Никита Андреевич	
	Рожок Софья Денисовна	
	Рюмин Матвей Михайлович	
Team 5	Сатышев Владислав Викторович	**RT 9**
	Свининников Иван Дмитриевич	
	Сивохина Анна Дмитриевна	
	Травников Иван Ильич	
	Фонарева Ксения Павловна	
Team 6	Хайруллин Алмаз Ильшатович	**RT 1**
	Ханин Вадим Александрович	
	Ходаковский Кирилл Антонович	
	Чеботарев Григорий Владимирович	
Team 7	Чернышева Анастасия Дмитриевна	**RT 10**
	Чечулин Николай Дмитриевич	
	Чуксеев Антон Ильич	
	Шашков Константин Александрович	

A. Khayrullin absent

Syllabus

- **Teams, Group 2**

Team 1	Абрамов Никита Сергеевич	**RT 8**
	Алмасян Санасар Багдасарович	
	Архипов Николай Алексеевич	
	Ашарин Игорь Максимович	
Team 2	Баринов Кирилл Алексеевич	**RT 1**
	Белоновский Пётр Ильич	
	Беляков Кирилл Олександрович	
	Борисов Артём Николаевич	
Team 3	Воронина Влада Александровна	**RT 11**
	Герцог Анна Андреевна	
	Голубкова Анна Ярославовна	
	Гончаров Антон Дмитриевич	
	(Григорьян Артём Юрьевич)	
	(Житний Григорий Дмириевич)	
Team 4	Захаренкова Елизавета Юрьевна	**RT 4**
	Иванов Артемий Андреевич	
	Казакова Елена Михайловна	
	Калинина Дарья Игоревна	
Team 5	Киричок Владислав Игоревич	**RT 2**
	Ковалёнок Иван Владимирович	
	Куликов Артём Валерьевич	
	Курлович Елизавета Юрьевна	
Team 6	Макаров Артём Максимович	**RT 3**
	Пак Александр Сергеевич	
	Смирнов Артем Денисович	
	Турчин Руслан Олегович	
Team 7	Тюпляев Никита Алексеевич	**RT 10**
	Якунин Иван Вадимович	
	Янковский Максим Олегович	

+ Nikolay Rukavishnikov

Syllabus: Research Topics to chose from for OralPres2

- **RT1. How does ChatGPT work? How could it be enhanced? Compare with the working/reasoning of the human brain.**
- **RT2. How does the human brain work? Are there fundamental brain/thinking mechanisms that could not be implemented on a computer?**
- **RT3. What is the difference between human intelligence and AI?**
- **RT4. What is the research done in top universities on the ethics of AI? Make a critical assessment. (Optional add-on: Which research projects could you propose?)**
- **RT5. What are key ethical issues related to the 'intelligence explosion' in AI spearheaded by openAI and ChatGPT? (The focus is here on the company openAI and its products.)**
- **RT6. Could computers have a mind and consciousness?**
- **RT7. Is thinking computing?**
- **RT8. Can ChatGPT think?**
- **RT9. How could AI evolve, ideally and less ideally?**
- **RT10. What are the existing/possible definitions of "artificial general intelligence" (AGI) and/or of "superintelligence"? Discuss the feasibility of superintelligence. (Use other sources than Bostrom!)**
- **RT11. What are some of the most important/interesting/surprising/risky business opportunities created by the 'intelligence explosion' in AI (e.g. ChatGPT) ? Give also a critical analysis from the ethical/societal point of view.**
- **RT12. What is XAI ? How could it be implemented in LLMs (Large Language Models) ?**

Philosophy of AI

General concepts and problems

Ch. 26, R&N

Philosophical questions (from Ch. 26, R&N)

- Main topic of chapter: what it means to think and whether artifacts could and should ever do so.

- **Weak AI** versus **strong AI**.

- AI that can act '*as if*' it can think (simulate) versus can *actually* think
- John Searle: strong AI = AI with a mind (consciousness) and mental states (real thinking can only be done by a mind).
- What is the position of most computer scientists, including R&N ?

Philosophical questions (from Ch. 26, R&N)

26.1 WEAK AI: CAN MACHINES ACT INTELLIGENTLY ?

- **Whether (weak) AI is possible depends on how it is defined.**
- **R&N defined AI as the quest for the best agent program on a given machine.**
- **According to such a definition, machines can indeed (of course) act intelligently.**

26.1 WEAK AI: CAN MACHINES ACT INTELLIGENTLY ?

- **But philosophers have asked a more deep-digging (demanding) question: can computers *really* think?**
- **The computer scientist Edsger Dijkstra (1984) said that "The question of whether *Machines Can Think* . . . is about as relevant as the question of whether *Submarines Can Swim*."**
- **Alan Turing, in his famous paper "Computing Machinery and Intelligence" (1950), suggested that instead of asking whether machines can think, we should ask whether machines can pass a behavioral intelligence test (Turing Test).**
- **5 min, 30% of time, by Y2000. R&N – not yet. ???**

26.1 WEAK AI: CAN MACHINES ACT INTELLIGENTLY ?

- **Turing examined a wide variety of possible objections to the possibility of intelligent machines (and weak AI).**

26.1.1 The argument from disability.

- **The "argument from disability" makes the claim that "a machine can never do X." As examples of X, Turing lists the following:**

- **Be kind, resourceful, beautiful, friendly, have initiative, have a sense of humor, tell right from wrong, make mistakes, fall in love, enjoy strawberries and cream, make someone fall in love with it, learn from experience, use words properly, be the subject of its own thought, have as much diversity of behavior as man, do something really new.**

26.1 WEAK AI: CAN MACHINES ACT INTELLIGENTLY ?

- Be kind, resourceful, beautiful, friendly, have initiative, have a sense of humor, tell right from wrong, make mistakes, fall in love [there has been only limited speculation about whether it is in fact likely (Kim et al., 2007)], enjoy strawberries and cream, make someone fall in love with it [teddybear], learn from experience, use words properly, be the subject of its own thought, have as much diversity of behavior as man, do something really new. RED: probably not possible for the time being.

- Chess, drive cars and helicopters, diagnose diseases, etc. etc., including small but significant discoveries in science.
- But algorithms also perform at human levels on tasks that seemingly involve human judgment, or as Turing put it, "learning from experience" and the ability to "tell right from wrong."

26.1 WEAK AI: CAN MACHINES ACT INTELLIGENTLY ?

- As far back as 1955, Paul Meehl studied the decision-making processes of trained experts at subjective tasks such as predicting the success of a student in a training program or the recidivism of a criminal.
- In 19 out of the 20 studies he looked at, Meehl found that simple statistical learning algorithms (such as linear regression or naive Bayes) predict better than the experts.

26.1.2 The mathematical objection (based on Goedel's Incompleteness Th.)

- GIT: Briefly, for any formal axiomatic system F that is consistent and powerful enough to do arithmetic, it is possible to construct a so-called Goedel sentence G(F) with the following properties:
  - G(F) is a sentence of F, but cannot be proved within F.
  - G(F) is true.

- GIT applies to the Turing machine (=universal computer (infinite)). So a Turing machine cannot compute everything; more precisely, it cannot prove the truth of all true mathematical statements.

26.1 WEAK AI: CAN MACHINES ACT INTELLIGENTLY ?

- **Philosophers such as J. R. Lucas (1961), following Goedel, have claimed that this theorem shows that machines are mentally inferior to humans, because machines are formal systems that are limited by the incompleteness theorem—they cannot establish the truth of their own Goedel sentence—while humans have no such limitation.**

- **This claim has caused decades of controversy, spawning a vast literature, including two books by the mathematician Sir Roger Penrose (1989, 1994) that repeat the claim with some fresh twists (such as the hypothesis that humans are different because their brains operate by quantum gravity (?!)).**
- **So Penrose claims that the brain is not like a *classical* computer, but more like a *quantum* computer, not restricted by GIT.**
- **Three problems with Lucas' claim:**

26.1 WEAK AI: CAN MACHINES ACT INTEL

Back-up

- **Counterargument 1) Computers are not really limited by GIT. If they were Turing machines, they would be. But they are finite and therefore only approximately Turing Machines.**

- **In R&N's words:** "Goedel's incompleteness theorem (GIT) applies only to formal systems that are powerful enough to do arithmetic. This includes Turing machines, and Lucas's claim is in part based on the assertion that computers are Turing machines. This is a good approximation, but is not quite true. Turing machines are infinite, whereas computers are finite, and any computer can therefore be described as a (very large) system in propositional logic, which is not subject to Goedel's incompleteness theorem."

26.1 WEAK AI: CAN MACHINES ACT INTEL

Back-up

- **Counterargument 2). EVEN IF a computer would be limited by GIT, that does not mean it cannot be intelligent. Humans are in the same situation. An agent should not be too ashamed that it cannot establish the truth of some sentence while other agents can.**
- **Consider the sentence**

  J. R. Lucas cannot consistently assert that this sentence is true.
- **If Lucas asserted this sentence, then he would be contradicting himself, so therefore Lucas cannot consistently assert it, and hence it must be true.**
- **We have thus demonstrated that there is a sentence that Lucas cannot consistently assert while other people (and machines) can. But that does not make us think less of Lucas.**

26.1 WEAK AI: CAN MACHINES ACT INTELLIGENTLY ?

- **Counterargument 3). Even if we grant that computers have limitations on what they can prove, there is no evidence that humans are immune from those limitations. (Repitition of argument 2.)**

- **R&N: "It is all too easy to show rigorously that a formal system cannot do X, and then claim that humans can do X using their own informal method, without giving any evidence for this claim. Indeed, it is impossible to prove that humans are not subject to G¨odel's incompleteness theorem, because any rigorous proof would require a formalization of the claimed unformalizable human talent, and hence refute itself. So we are left with an appeal to intuition that humans can somehow perform superhuman feats of mathematical insight. This appeal is expressed with arguments such as "we must assume our own consistency, if thought is to be possible at all" (Lucas, 1976). But if anything, humans are known to be inconsistent."**

26.1 WEAK AI: CAN MACHINES ACT INTELLIGENTLY ?

26.1.3 The argument from informality of behaviour.

- **This is the claim that human behavior is far too complex to be captured by any simple set of rules and that because computers can do no more than follow a set of rules, they cannot generate behavior as intelligent as that of humans.**

- **The principal proponent of this view has been the philosopher Hubert Dreyfus: What Computers Can't Do (1972), the sequel What Computers Still Can't Do (1992), and, with his brother Stuart, Mind Over Machine (1986).**

26.1 WEAK AI: CAN MACHINES ACT INTELLIGENTLY ?

- **R&N claim that Dreyfus' arguments only concern GOFAI (Good Old-Fashioned AI, i.e. based on first-order logical rules without learning), so not more advanced AI as ML (ANN).**
- **Dreyfus: For instance in social contexts, one apparently has "a direct sense of how things are done and what to expect [e.g. give present]". We do not always act by consciously following rules. Other example: intuitive chess playing.**

- **R&N: "It is certainly true that much of the thought processes of a present-giver or grandmaster is done at a level that is not open to introspection by the conscious mind. But that does not mean that the thought processes do not exist. The important question that Dreyfus does not answer is *how* the right move gets into the grandmaster's head."**
- **So the debate on whether AI can act intelligently goes on!**

26.2 STRONG AI: CAN MACHINES REALLY THINK?

26.2 STRONG AI: CAN MACHINES REALLY THINK?

- **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.**
- **Again, the objection was foreseen by Turing. He cites a speech by Professor Geoffrey Jefferson (1949):**
  - **"Not until a machine could write a sonnet or compose a concerto because of thoughts and emotions felt, and not by the chance fall of symbols, could we agree that machine equals brain—that is, not only write it but know that it had written it."**
- **Typically, philosophers who believe that machines can (never) really think, argue that machines have no 'PHENOMENOLOGY (FEELINGS, EMOTIONS)' and (therefore) no CONSCIOUSNESS (AWARENESS).**

26.2 STRONG AI: CAN MACHINES REALLY THINK ?

- Turing maintains that the question is just as ill-defined as asking, “Can machines behave intelligently?” (the question of weak AI).
- Why should we insist on a higher standard for machines than we do for humans?
- Turing’s “polite convention”

- R&N believe that in the (near) future we will make no distinction between “real” and “artificial” thinking… because we will get used to very intelligent robots.

- Philosophers who believe that strong AI is possible (in principle) use this argument: mental processes are physical-chemical brain processes → mental states are physical states → physical states can be reproduced by advanced technology → strong AI seems possible in principle.

Philosophy and Ethics of AI

Louis Vervoort, PhD

Higher School of Economics
September-October 2023

Syllabus

- **Schedule Book chapter presentations (OralTest1) and Research presentations (OralTest2)**

	Lectures (Tuesdays)	Book chapter pres (Seminar, Thursday)	Research presentation (Seminar, Thursday)
Week 1, Tue 5 Sept	1, 2	- - -	- - -
Week 2, Tue 12 Sept	3	Team 1, Chapter 1	- - -
Week 3, Tue 19 Sept	4	NO SEMINAR	NO SEMINAR
Week 4, Tue 26 Sept	5	Team 2, Chapter 2 WED 27 ! Team 3, Chapter 3 THUR 28 !	Team 7 WED 27 sept ! Team 6 THUR 28 sept !
Week 5, Tue 3 Oct	6 + Test1	Team 4, Chapter 4	Team 5
Week 6, Tue 10 Oct	7	Team 5, Chapter 5	Team 4
Week 7, Tue 17 Oct	8	Team 6, Chapter 6 Wed. Oct 18 Team 7, Chapter 7 Thur Oct 19	Team 3 Wed. Oct. 18 Team 2 Thur. Oct. 19.
Week 8, Tue 24 Oct	Half lect/sem.	- - -	Team 1 (on Tue 24/10, during lecture)
Week 9, Tue 31 Oct	Exam1	- - -	- - -

Syllabus

- **Teams, Group 1**

Team 1	(Асташкин Артемий Андреевич) (absent)	
	Гнездилова Вера Андреевна	**RT 5**
	Дмитриченко Камила Дмитриевна	
	Досаев Савелий Юрьевич	
Team 2	Камбачеков Тимур Алимович	**RT 3**
	Кульжик Степан Михайлович	
	Куракина Анеля Романовна	
	Лапко Дарья Андреевна	
Team 3	Луговцев Тимур Андреевич	**RT 2**
	Малышева Мария Александровна	
	Манякин Павел Дмитриевич	
	Мирзак Александр Сергеевич	
	Нигматуллина Диана Дмитриевна	
Team 4	Ноздрина Марина Викторовна	**RT 11**
	Петров Илья Родионович	
	Раташнюк Никита Андреевич	
	Рожок Софья Денисовна	
	Рюмин Матвей Михайлович	
Team 5	Сатышев Владислав Викторович	**RT 9**
	Свининников Иван Дмитриевич	
	Сивохина Анна Дмитриевна	
	Травников Иван Ильич	
	Фонарева Ксения Павловна	
Team 6	Хайруллин Алмаз Ильшатович	**RT 1**
	Ханин Вадим Александрович	
	Ходаковский Кирилл Антонович	
	Чеботарев Григорий Владимирович	
Team 7	Чернышева Анастасия Дмитриевна	**RT 10**
	Чечулин Николай Дмитриевич	
	Чуксеев Антон Ильич	
	Шашков Константин Александрович	

A. Khayrullin absent

Syllabus

- **Teams, Group 2**

Team 1	Абрамов Никита Сергеевич	**RT 8**
	Алмасян Санасар Багдасарович	
	Архипов Николай Алексеевич	
	Ашарин Игорь Максимович	
Team 2	Баринов Кирилл Алексеевич	**RT 1**
	Белоновский Пётр Ильич	
	Беляков Кирилл Олександрович	
	Борисов Артём Николаевич	
Team 3	Воронина Влада Александровна	**RT 11**
	Герцог Анна Андреевна	
	Голубкова Анна Ярославовна	
	Гончаров Антон Дмитриевич	
	(Григорьян Артём Юрьевич)	
	(Житний Григорий Дмириевич)	
Team 4	Захаренкова Елизавета Юрьевна	**RT 4**
	Иванов Артемий Андреевич	
	Казакова Елена Михайловна	
	Калинина Дарья Игоревна	
Team 5	Киричок Владислав Игоревич	**RT 2**
	Коваленок Иван Владимирович	
	Куликов Артём Валерьевич	
	Курлович Елизавета Юрьевна	
Team 6	Макаров Артём Максимович	**RT 3**
	Пак Александр Сергеевич	
	Смирнов Артем Денисович	
	Турчин Руслан Олегович	
Team 7	Тюпляев Никита Алексеевич	**RT 10**
	Якунин Иван Вадимович	
	Янковский Максим Олегович	

+ Nikolay Rukavishnikov

Syllabus: Research Topics to chose from for OralPres2

- RT1. How does ChatGPT work? How could it be enhanced? Compare with the working/reasoning of the human brain.
- RT2. How does the human brain work? Are there fundamental brain/thinking mechanisms that could not be implemented on a computer?
- RT3. What is the difference between human intelligence and AI?
- RT4. What is the research done in top universities on the ethics of AI? Make a critical assessment. (Optional add-on: Which research projects could you propose?)
- RT5. What are key ethical issues related to the 'intelligence explosion' in AI spearheaded by openAI and ChatGPT? (The focus is here on the company openAI and its products.)
- RT6. Could computers have a mind and consciousness?
- RT7. Is thinking computing?
- RT8. Can ChatGPT think?
- RT9. How could AI evolve, ideally and less ideally?
- RT10. What are the existing/possible definitions of "artificial general intelligence" (AGI) and/or of "superintelligence"? Discuss the feasibility of superintelligence. (Use other sources than Bostrom!)
- RT11. What are some of the most important/interesting/surprising/risky business opportunities created by the 'intelligence explosion' in AI (e.g. ChatGPT) ? Give also a critical analysis from the ethical/societal point of view.
- RT12. What is XAI ? How could it be implemented in LLMs (Large Language Models) ?

What is Ethics ? (1h)

Ethics

- **Text "Ethics", James Fieser (Internet Encyc. Phil.):**

- **Philosophers today usually divide ethical theories into three general subject areas: metaethics, normative ethics, and applied ethics.**

- **Meta-ethics investigates where our ethical principles come from, and what they mean. Are they merely social inventions? Do they involve more than expressions of our individual emotions? Metaethical answers to these questions focus on the issues of universal truths, the will of God, the role of reason in ethical judgments, and the meaning of ethical terms themselves.**

Ethics

- **Normative ethics takes on a more practical task, which is to arrive at moral standards that regulate right and wrong conduct.**
- **This may involve articulating the good habits that we should acquire, the duties that we should follow, or the consequences of our behavior on others.**

- **Applied Ethics**

Meta-ethics

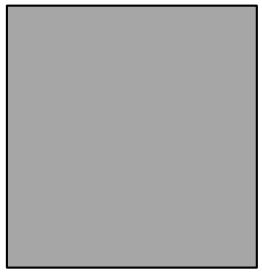
ii. Emotion and Reason

A second area of moral psychology involves a dispute concerning the role of reason in motivating moral actions. If, for example, I make the statement "abortion is morally wrong," am I making a rational assessment or only expressing my feelings? On the one side of the dispute, 18th century British philosopher David Hume argued that moral assessments involve our emotions, and not our reason. We can amass all the reasons we want, but that alone will not constitute a moral assessment. We need a distinctly emotional reaction in order to make a moral pronouncement. Reason might be of service in giving us the relevant data, but, in Hume's words, "reason is, and ought to be, the slave of the passions." Inspired by Hume's anti-rationalist views, some 20th century philosophers, most notably A.J. Ayer, similarly denied that moral assessments are factual descriptions. For example, although the statement "it is good to donate to charity" may on the surface look as though it is a factual description about charity, it is not. Instead, a moral utterance like this involves two things. First, I (the speaker) I am expressing my personal feelings of approval about charitable donations and I am in essence saying "Hooray for charity!" This is called the *emotive* element insofar as I am expressing my emotions about some specific behavior. Second, I (the speaker) am trying to get you to donate to charity and am essentially giving the command, "Donate to charity!" This is called the *prescriptive* element in the sense that I am prescribing some specific behavior.

Metaethics

From Hume's day forward, more rationally-minded philosophers have opposed these emotive theories of ethics (see non-cognitivism in ethics) and instead argued that moral assessments are indeed acts of reason. 18th century German philosopher Immanuel Kant is a case in point. Although emotional factors often do influence our conduct, he argued, we should nevertheless resist that kind of sway. Instead, true moral action is motivated only by reason when it is free from emotions and desires. A recent rationalist approach, offered by Kurt Baier (1958), was proposed in direct opposition to the emotivist and prescriptivist theories of Ayer and others. Baier focuses more broadly on the reasoning and argumentation process that takes place when making moral choices. All of our moral choices are, or at least can be, backed by some reason or justification. If I claim that it is wrong to steal someone's car, then I should be able to justify my claim with some kind of argument. For example, I could argue that stealing Smith's car is wrong since this would upset her, violate her ownership rights, or put the thief at risk of getting caught. According to Baier, then, proper moral decision making involves giving the best reasons in support of one course of action versus another.

Ethics



- **7 deadly sins:**

- **pride, greed, wrath (= anger, rage), envy, lust, gluttony and sloth (= laziness)**

Ethics

- **Pride (Latin: superbia) is considered, on almost every list, the original and most serious of the seven deadly sins. Out of the seven, it is the most angelical, or demonic.[40] It is also thought to be the source of the other capital sins. Also known as hubris (from ancient Greek ὕβρις), or futility, it is identified as dangerously corrupt selfishness, the putting of one's own desires, urges, wants, and whims before the welfare of other people.**

- **In even more destructive cases, it is irrationally believing that one is essentially and necessarily better, superior, or more important than others, failing to acknowledge the accomplishments of others, and excessive admiration of the personal image or self (especially forgetting one's own lack of divinity, and refusing to acknowledge one's own limits, faults, or wrongs as a human being).**

> **What the weak head with strongest bias rules, Is pride, the never-failing vice of fools.**
>
> **— Alexander Pope, An Essay on Criticism, line 203.**

- **As pride has been labelled the father of all sins, it has been deemed the devil's most prominent trait. C.S. Lewis writes, in Mere Christianity, that pride is the "anti-God" state, the position in which the ego and the self are directly opposed to God: "Unchastity, anger, greed, drunkenness, and all that, are mere fleabites in comparison: it was through Pride that the devil became the devil: Pride leads to every other vice: it is the complete anti-God state of mind."[41] Pride is understood to sever the spirit from God, as well as His life-and-grace-giving Presence.[5]**

2. Normative Ethics

- **Virtue theories (p. 6-)**
- **Duty theories (p. 7-)**
- **Consequentialist theories (p. 9-)**
- **Group 4: Types of utilitarianism & Ethical egoism (p. 10-)**
  - **E.g. social contract theory, which is a type of rule-ethical-egoism.**

Ethics

- **I. Virtue theories: to act ethically is to act according to certain key virtues**
  - **benevolence [доброжелательность, желать = to wish]**
  - **Plato: cardinal virtues: wisdom, courage, temperance and justice.**
  - **Other important virtues are fortitude, generosity, self-respect, good temper, and sincerity.**
  - **theological virtues: faith, hope, and charity.**

- **Aristotle argued that virtues are good habits that we acquire, which regulate our emotions.**

- **II. Duty theories base morality on specific, foundational principles of duty, of obligation. ("deontological" theories)**
- **"nonconsequentialist" since these principles are obligatory, irrespective of the consequences that might follow from our actions.**

- **1. Pufendorf: duties towards God, duties towards oneself (e.g. developing one's skills and talents, not harming one's body,…), duties towards others**

- **2. A second duty-based approach to ethics is rights theory.**
  - **Rights and duties are correlated.**
  - **Locke: the laws of nature mandate that we should not harm anyone's life, health, liberty or possessions: natural rights.**
  - **United States Declaration of Independence (Thomas Jefferson) recognizes three foundational rights: life, liberty, and the pursuit of happiness. These induce other moral rights, including the rights of property, movement, speech, and religious expression.**
  - **Locke and Jefferson et al. considered these as natural (= not imposed by government), universal (valid in all countries), equal for all people alike.**

Ethics

- **3. A third duty-based theory is that by Kant, which emphasizes a single principle of duty.**
- **Single, self-evident principle of reason that he calls the "categorical imperative":**
  - **Treat people as an end, and never as a means to an end. Always treat people with dignity, and never use them as mere instruments.**
  - **Act only according to that maxim whereby you can will that it should become a universal law.**

- **III. Consequentialist Theories: An action is morally right if the consequences of that action are more favorable than unfavorable.**
  - **Problem: favorable for who? Presumably, for "everyone", or at least "as many people as possible" (utilitarianism)**
  - **Most famous consequentialist theory is utilitarianism (Bentham, Mill): "an ethical deed is one that maximizes happiness & wellbeing for a majority of people".**
- **Teleological theories (from the Greek word telos, or end) since the end result of the action is the sole determining factor of its morality.**

Philosophy of AI

General concepts and problems

Ch. 26, R&N

26.2 STRONG AI: CAN MACHINES REALLY THINK?

26.2 STRONG AI: CAN MACHINES REALLY THINK?

- **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.**
- **Again, the objection was foreseen by Turing. He cites a speech by Professor Geoffrey Jefferson (1949):**
  - **"Not until a machine could write a sonnet or compose a concerto because of thoughts and emotions felt, and not by the chance fall of symbols, could we agree that machine equals brain—that is, not only write it but know that it had written it."**
- **Typically, philosophers who believe that machines can (never) really think, argue that machines have no 'PHENOMENOLOGY (FEELINGS, EMOTIONS)' and (therefore) no CONSCIOUSNESS (AWARENESS).**

26.2 STRONG AI: CAN MACHINES REALLY THINK ?

- Turing maintains that the question is just as ill-defined as asking, “Can machines behave intelligently?” (the question of weak AI).
- Why should we insist on a higher standard for machines than we do for humans?
- Turing’s “polite convention”

- R&N believe that in the (near) future we will make no distinction between “real” and “artificial” thinking… because we will get used to very intelligent robots.

- Philosophers who believe that strong AI is possible (in principle) use this argument: mental processes are physical-chemical brain processes → mental states are physical states → physical states can be reproduced by advanced technology → strong AI seems possible in principle.

26.2 STRONG AI: CAN MACHINES REALLY THINK ?

Searle's argument against strong AI. Biological naturalism and the Chinese Room Argument (CRA).

- **The question whether strong AI is possible can be studied from this angle: can mind/mental states/understanding be created/had by a computer, or can only biological neurons (a very specific biological-chemical-physical system) generate/have them ? → CRA:**

- **Searle describes a hypothetical "system" that is clearly running a program and passes the Turing Test, but that equally clearly (according to Searle) does not understand anything of its inputs and outputs.**
- **His conclusion is that running the appropriate program (i.e., having the right outputs) is not a sufficient condition for being a mind.**
- **In a slogan: according to Searle, a computer or AI cannot *really* think.**

26.2 STRONG AI: CAN MACHINES REALLY THINK ?

Searle's argument against strong AI. Biological naturalism and the Chinese Room Argument (CRA).

- **CRA, from R&N: The "system" consists of a human, who understands only English, equipped with a rule book, written in English, and various stacks of paper, some blank, some with indecipherable Chinese inscriptions. (The human therefore plays the role of the CPU, the rule book is the program, and the stacks of paper are the storage device.) The system is inside a room with a small opening to the outside. Through the opening appear slips of paper with indecipherable symbols. The human finds matching symbols in the rule book, and follows the instructions. The instructions may include writing symbols on new slips of paper, finding symbols in the stacks, rearranging the stacks, and so on. Eventually, the instructions will cause one or more symbols to be transcribed onto a piece of paper that is passed back to the outside world.**

26.2 STRONG AI: CAN MACHINES REALLY THINK ?

- From the outside, we see a system that is taking input in the form of Chinese sentences and generating answers in Chinese that are as "intelligent" as those in the conversation imagined by Turing. This 'Chinese Room' could pass a Turing test. Searle then argues: the person in the room does not understand Chinese (given). The rule book and the stacks of paper, being just pieces of paper, do not understand Chinese. Therefore, there is no understanding of Chinese.
- **Hence, according to Searle, running the right program does not necessarily generate understanding.**
- By analogy, Searle argues, computers running programs do not have understanding, just as the human in the Chinese Room does not have understanding of Chinese – he is just following rules.
- **Or: computers cannot have minds.**

26.2 STRONG AI: CAN MACHINES REALLY THINK ?

- **Hence, according to Searle, running the right program does not necessarily generate understanding.**
- **Various objections have been proposed against the CRA, including by R&N.**
- **(Of course, the CRA is not a proof, just an argument.)**
- **Some say that in the CRA indeed the program (the rule book) on its own does not have understanding, but the total system (the CR) has. (In analogy, individual neurons have no understanding, but the whole brain has, one might say.) The human brain can / should be compared to the whole CR. So this counterargument against the CRA aims at arguing that a complex system (program + CPU + memory essentially) CAN have understanding.**

26.2 STRONG AI: CAN MACHINES REALLY T *Back-up*

- Searle in later texts described his argument in more detail, starting from 3 premises / axioms:
  1. Syntax is not sufficient for semantics.
  2. Computer programs are entirely defined by their formal, or syntactical, structure.
  3. Minds have mental contents; specifically, they have semantic contents.
- **From this he concludes: No computer program by itself is sufficient to give a system a mind. Programs, in short, are not minds, and they are not by themselves sufficient for having minds.**
- **What do you think yourselves ?**

- LV: I can agree with the idea that programs (algos) alone in the narrow sense (GOFAI) do not generate/have understanding. But we can enhance AI with ANN and with (background) knowledge-bases. I believe this is sufficient for rational consciousness / understanding / real thinking. Reason: such enhanced AI can master th-based thinking.
- So I think the weak axiom is axiom 2.

26.2 STRONG AI: CAN MACHINES REALLY THINK ?

- **Biological naturalism:** Searle believes that only minds have real understanding, and that minds obtain this understanding from the neurons :
- According to biological naturalism 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.
- A famous quote by Searle: brains cause minds.
- So according to Searle, only very specific biochemical structures (neurons) can generate minds & understanding.
- R&N: Thus, according to Searle, mental states cannot be duplicated just on the basis of some program having the same functional structure with the same input–output behavior; we would require that the program be running on an architecture with the same causal power as neurons.