

# Philosophy of AI and Consciousness

**Philosophical perspectives on the exact sciences I**

EPFL, Oct 4, 2023

Alin Christoph Cucu

# Three questions

- Could AI become conscious?
- Is AGI possible? (« Could AI think? »)
- Is the human mind a computer?

# I. Consciousness, Thinking, Selfhood and AI

ARTIFICIAL INTELLIGENCE

# What is consciousness? ChatGPT and advanced AI might redefine our answer

Technologists broadly agree that AI chatbots are not self-aware just yet, but there is some thought that we may have to re-evaluate how we talk about sentience.



Health Space Physics Technology Environment Mind Humans Life Mathematics Chemistry Earth Society

## Technology

# Can AI ever become conscious and how would we know if that happens?

It sounds far-fetched, but researchers are trying to recreate subjective experience in AIs, even if disagreement over what consciousness is will make it difficult to test

By [Thomas Lewton](#)

📅 25 July 2023



New Technology

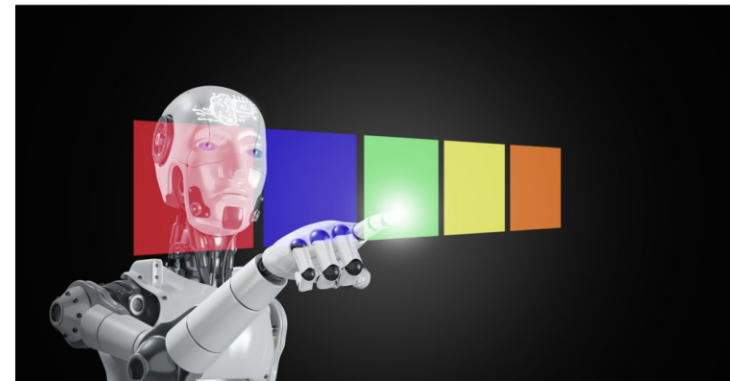
# Uh Oh, Chatbots Are Getting a Teeny Bit Sentient

Who could have seen that coming?



BY [TIM NEWCOMB](#) PUBLISHED: APR 17, 2023

🔖 SAVE ARTICLE



# Knowledge, not opinions

- **Opinion** = unjustified belief. Insubstantial like the images of physical things
- **Knowledge** = justified (true) belief. Ideally solid and unchanging like Platonic forms



# Opinions

- Ilya Sutskever, a co-founder of OpenAI, the company behind ChatGPT, has **speculated** that the algorithms behind his company's creations might be “**slightly conscious.**”
- David Chalmers: while ChatGPT doesn't clearly possess a lot of commonly assumed elements of consciousness, like sensation and independent agency, it's **easy to imagine** that a more sophisticated program could.
- Oxford philosopher Nick Bostrom believes AI has already started to show **small amounts of sentience.**

# Opinions

- "If you admit that it's **not an all-or-nothing** thing, then it's not so dramatic to say that some of these [AI] assistants might plausibly be candidates for having some degrees of sentience."
- "they exhibit **glimpses of creativity, insight, and understanding** that are quite impressive and may show the rudiments of reasoning."
- "It might not be practical to give it everything at once. I mean, I'd like to have a billion dollars. But if there are really trivial things that we could give them, like just changing a little thing in the code, that might matter a lot."



# The quest for KNOWLEDGE



# Consciousness (Nagel 1980, Chalmers 1996)

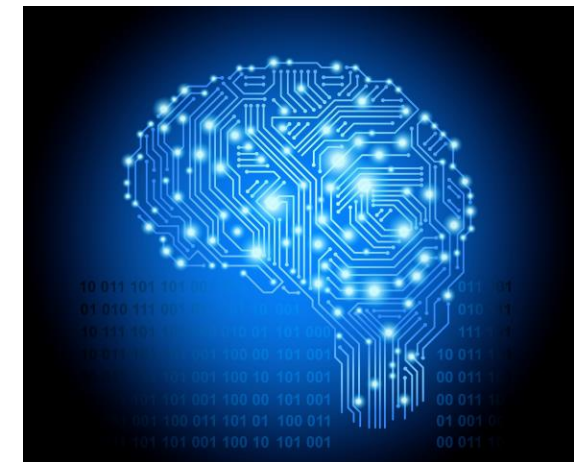
## Phenomenal consciousness

- Experience
- “What it’s like to be...”



## Psychological/access consciousness

- Explicable purely mechanically (in terms of computations implemented in matter)
- Explains behavior



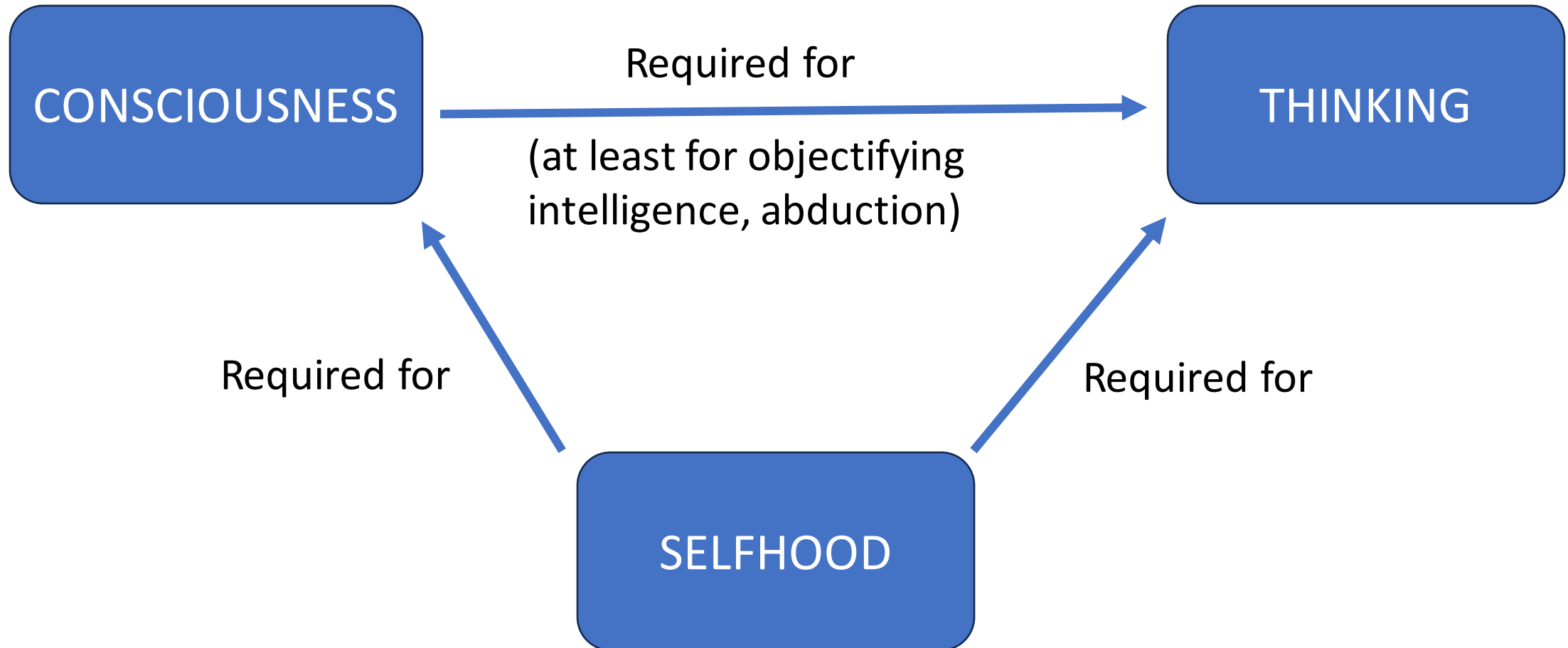
(Human) thinking

- **General** problem solving ability
- **Objectifying** intelligence: involves intentionality/aboutness. There are **objects of thought**, concepts, categories
- Distinct from **primal intelligence** which can be found also in animals (instinct-, disposition-based)
- **Deduction, induction** and **abduction** (coming up with hypotheses; tied to **imaginative** or **creative** thinking)

# Selfhood

- We can **distinguish** between ourselves and our thoughts/feelings etc.
- It is a self that **has** thoughts, feelings etc.
- Whether or not one can actually perceive the self (pro: Avicenna, Descartes, contra: Hume): it must **necessarily** be there

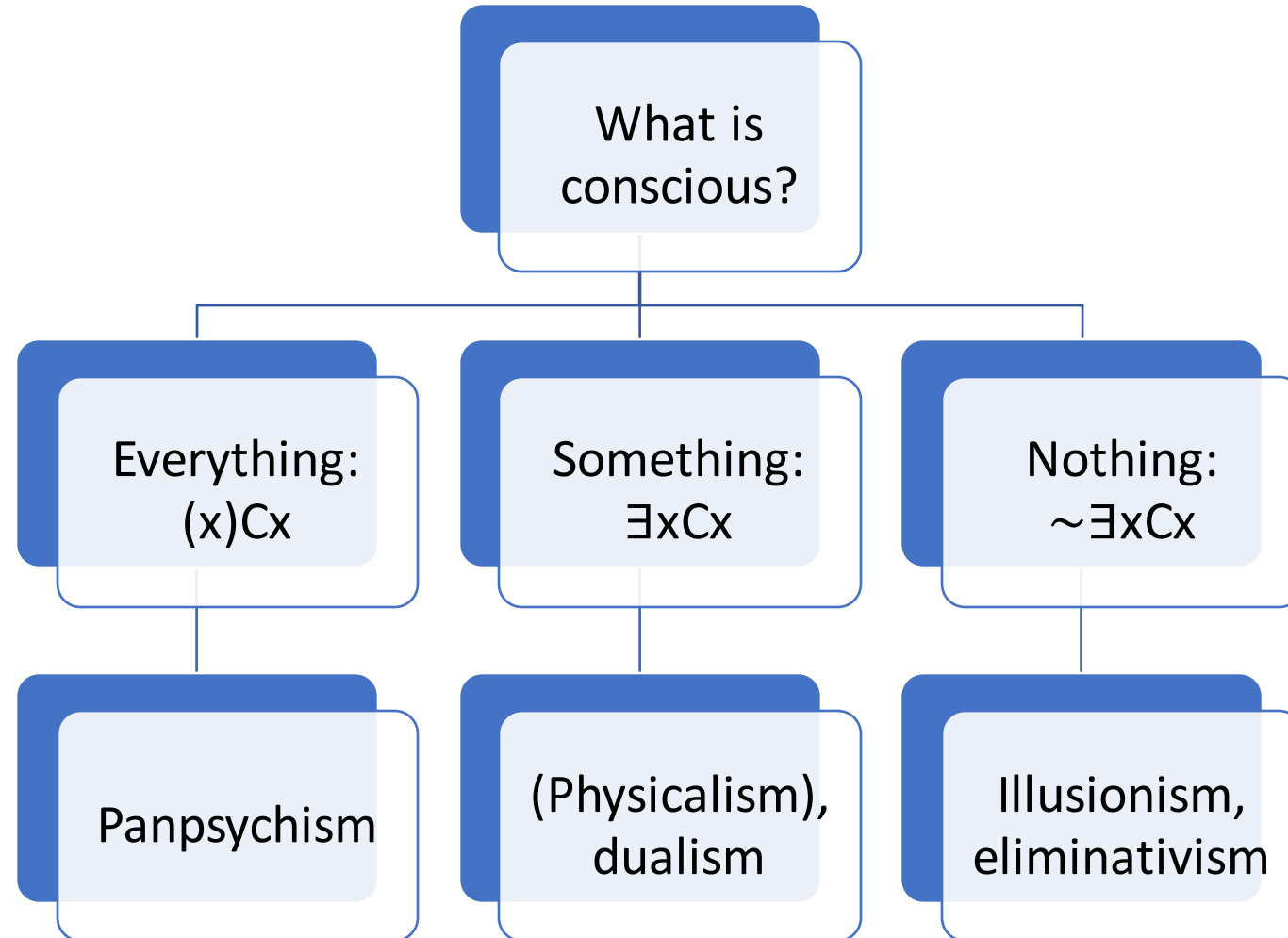
# Overview



# Doubts about Chalmers' approach

- Complete **neglect of the self** as the metaphysical prerequisite for a mental life
- Access consciousness not rather **tied** to phenomenal consciousness?
- **Behavior** really explainable without reference to experience?
- Cognition may **not be computational** and even be **immaterial** (see section III.)

# Global options for consciousness



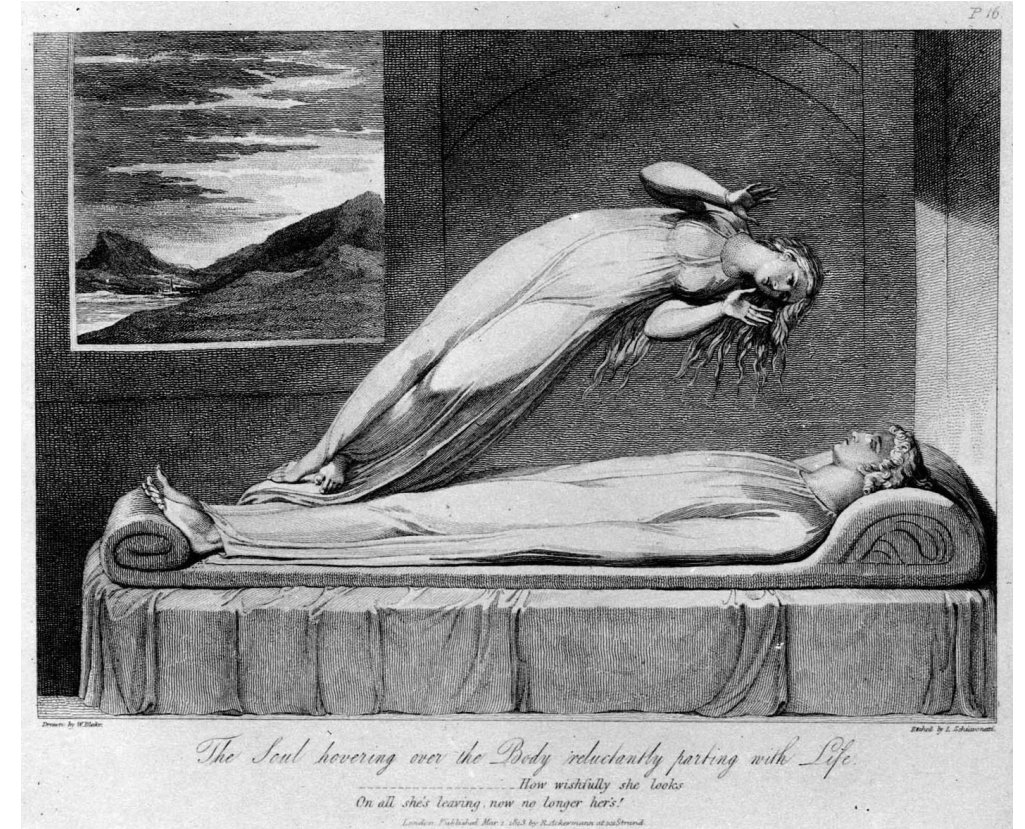
# Physicalism

- **Mental states are identical to physical states.**
- There is no « non-reductive physicalism » (see Kim 1998)
- Varieties mainly in function of whether identity is a *token* or *type* identity
- Seems not to **explain** consciousness but to explain it **away**
- On a physicalist picture, behavioral signs suffice



# Dualism

- Classically (as early as Plato): **Substance** dualism (soul = person = immaterial, mental substance)
  - Explains consciousness, self and reason in a straightforward way
- Thomistic **hylomorphism** (soul = form; form + matter = person)
- Main objection: interaction



Schiavonetti: *The Soul Leaving the Body* (1808)



# Chalmers' property dualism

- **Zombie argument:** if it is conceivable that there be an exact physical duplicate of me but without phenomenal consciousness, then phenomenal consciousness is irreducibly **non-physical**
- Problems: only mental **properties**, no account of the **self**, epiphenomenalism (**no freedom**)



# Panpsychism

- View that at the **fundamental** physical level, nature is mental (conscious and/or cognitive)
- Main problem: **combination problem** (how do many little minds/selves combine into unified minds/selves?)
- Runs counter to our intuitions
- No independent evidence
- Avoiding the pitfalls of physicalism without commitment to souls



Philipp Goff

# Could AI become conscious?

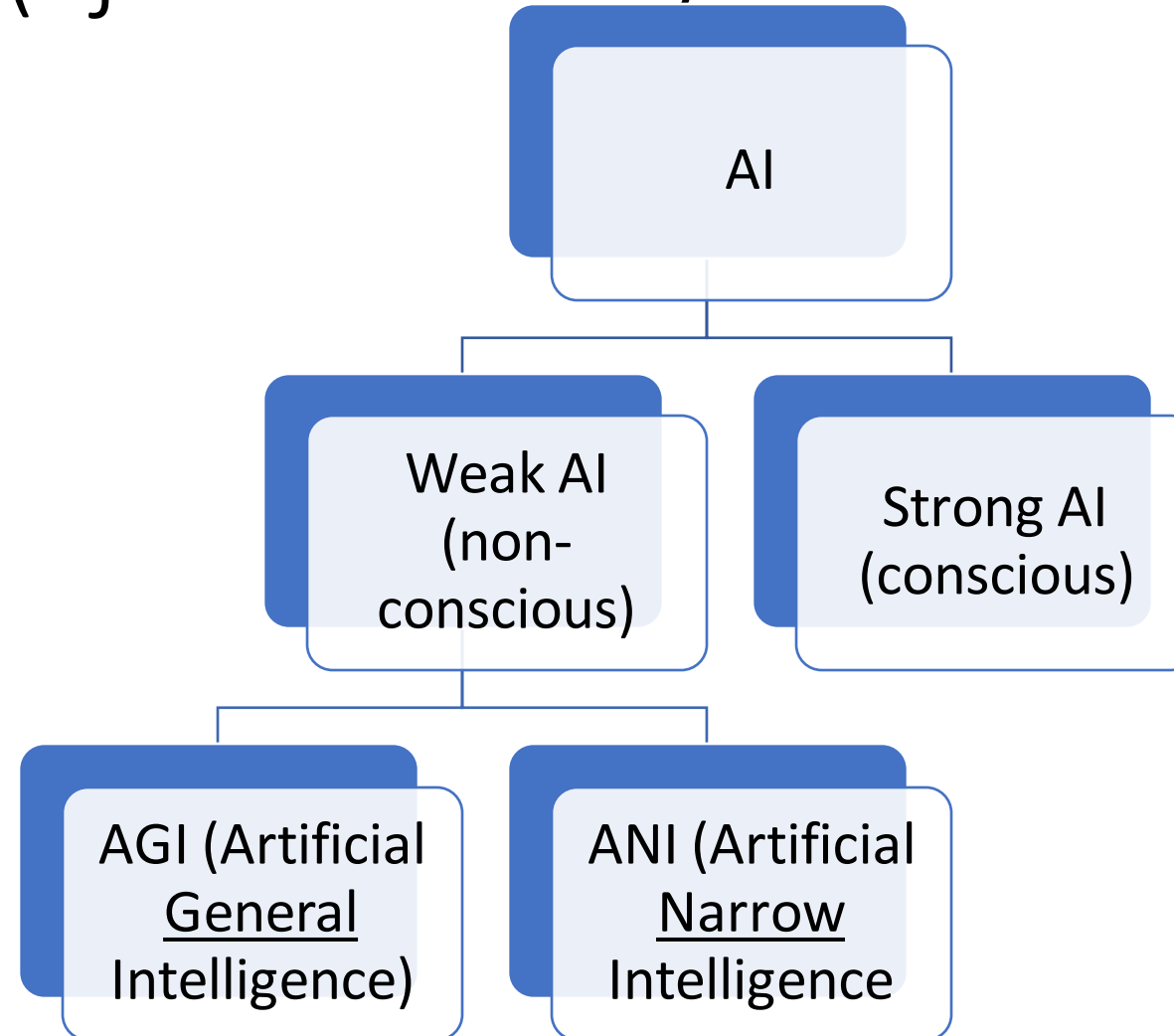
- Ontologies on which this is possible:
  - Property dualism
  - Panpsychism
  - (Emergent substance dualism)
  - Physicalism (with consciousness redefined)
- Central question: which physical structures or processes give rise to consciousness?
  - Computation: but Maudlin (1989) strongly objects
  - Integrated Information: Chalmers' favorite
  - Quantum processes: Penrose/Hameroff

# II. Is Artificial General Intelligence Possible?

# Framing the question

- Why not « Could AI / computers **think?** »
- Because thinking arguably comes with consciousness and a self (see above)
- Better to stick to **general problem solving capabilities** (hard enough!)

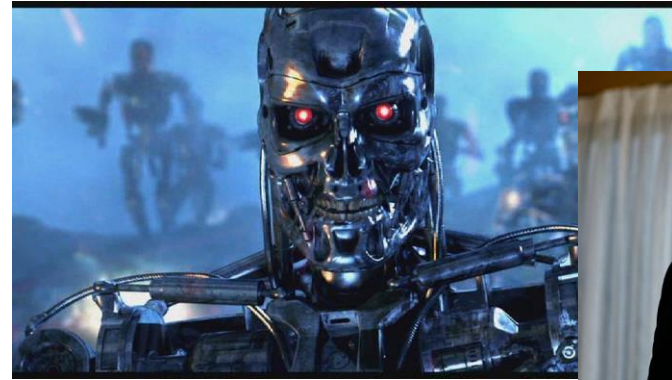
# Kinds of AI (Fjelland 2020)





# Strong AI

- Full range of human mental capacities (self-consciousness, desires, perhaps morality, perhaps wisdom...)
- E.g. Skynet, Ultron, A.I.D.A., R2-D2
- Purely fictional

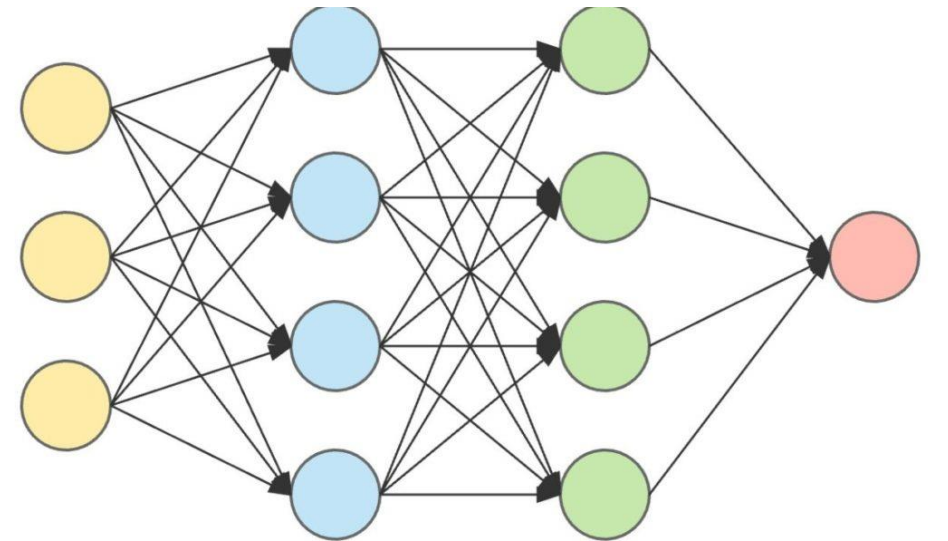
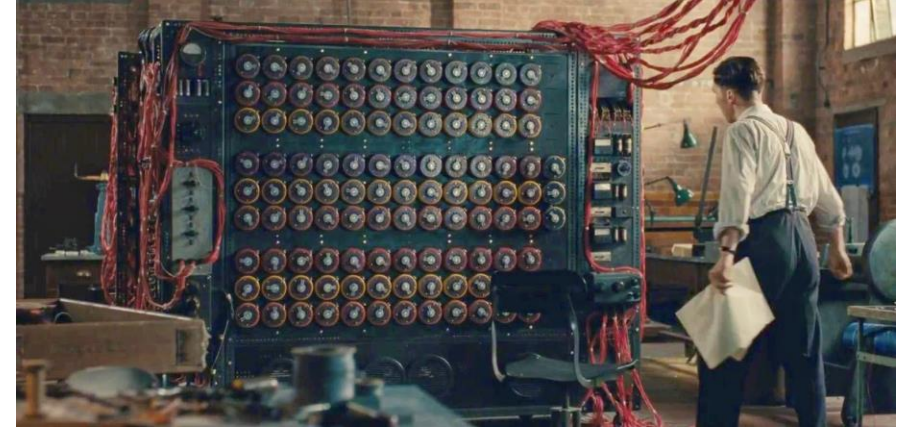


If AGI is possible, then strong AI *might* be possible; if not, then *a fortiori* strong AI is impossible.



# Implementation types (of ANI)

- Purely symbolic, deterministic, deductive: **Turing machines (TM)**
- Connectionist, probabilistic, inductive: **Neural networks (NN)**
- But NN can be implemented by universal Turing machines (Graves et al. 2014)



input layer

hidden layer 1

hidden layer 2

output layer

AGI is possible (Chalmers 2010)

1. The human brain is a **machine**.
2. We will be able to **emulate** this machine.
3. If we emulate this machine, there will be AGI.
4. Absent defeaters, there will be AGI.

AGI is possible (Chalmers 2010)

1. The human brain is a **machine**.

It's not, it's a **complex system** (Landgrebe & Smith 2023)

2. We will be able to **emulate** this machine.

We won't, because complex systems are **impossible to model mathematically** (Landgrebe & Smith 2023)

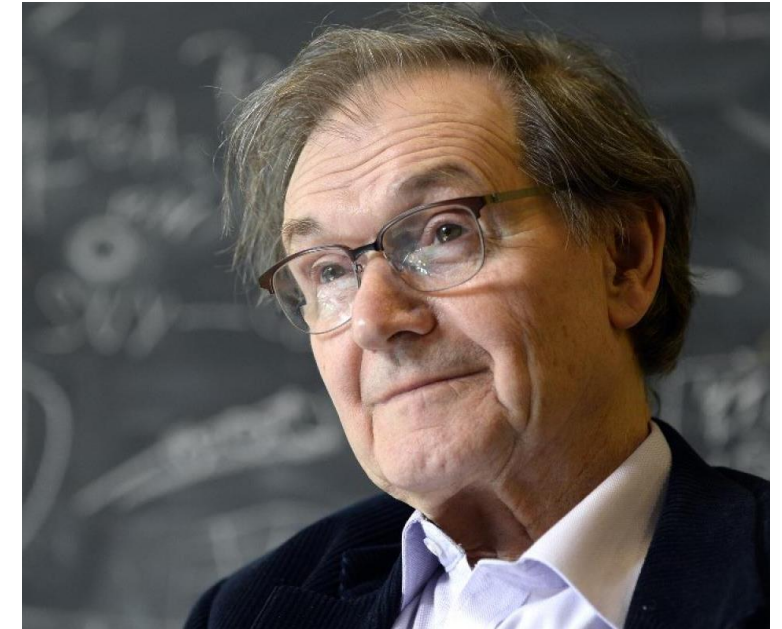
3. If we emulate this machine, there will be AGI.

What if **substance dualism** is true?

4. Absent defeaters, there will be AGI.

# Arguments against AGI

- **Gödel-style** refutations (Lucas 1961, Penrose 1995)
- **Embodiment**, « being-in-the-world » (Dreyfus 1972, Dreyfus & Dreyfus 1986)
- **The framing problem** (Fodor 1987) – see section III
- **The immateriality of thought** (Ross 1992, Feser 2013) – see section III



Sir Roger Penrose

# The Dreyfus argument

- Dreyfus (1967): human cognition is not « third-person » processing of atomistic, rule-connected data
- But depends on « embodiment » (« in-the-worldness »)
- Pattern recognition: holistic, parts get their meaning from the whole: « We posit meaning and infer objects »
- Problem solving: we pick out relevant data (see also framing problem)
- Also, humans gain knowledge by **manipulating** objects



# The Dreyfus argument, revised

- It will not do to put an AI into a robot « body »!
- E.g. tool use does not consist in calculations in physical space
- « Embodiment » should rather read « **embodied consciousness** » or « **embodied perspective** »
- Progress from TM to NN makes no difference
  - NN may have « tacit knowledge » but none of the other traits of human thought)
  - Also see TM implementability of NN





# Neural networks

- Neural networks (NN): very good at **pattern recognition**, can learn w/o explicit instructions
- Taken to have refuted Dreyfus' argument
  - But implementability of NN by TM
  - Still no holistic recognition (small disturbances lead to breakdown)
  - Also: variables intransparent. Cf. 'snow detector' case



# Big Data: end of causal thinking?

- ‘Data deluge’ makes causal thinking obsolete (only correlations can be found)
- Mayer-Schönberger & Cukier (2014): We cannot construct strong AI, but we *should align* our way of thinking with a correlational Big Data approach





# Ditch causality?

- But we cannot dispose of causality because of temporal order
  - Correlation is a symmetric relation: if A is correlated with B, then B is correlated with A
  - Causality is asymmetric: if C is the cause of E, E is not the cause of C
  - Things can be correlated but not causally related
- We know from our own agency what causality is (Swinburne 2005)

# To be fair: ChatGPT (2023)

Alin: If I shake your hand, whose hand am I holding?

ChatGPT: If you were to shake my hand, you would be holding the hand of the person you are interacting with in real life. However, it's important to note that I am a text-based AI and do not have a physical presence or a hand to shake. My responses are generated through text, and I don't possess a physical body. So, in a virtual sense, **you wouldn't be holding anyone's hand** when interacting with me.

A: Do you know what you just said or did you glean your answer from myriads of similar conversations on the internet?

C: **I generate responses based on patterns and information present in the text data I was trained on, up to my last knowledge update in September 2021. I don't possess consciousness or awareness, so I don't "know" things in the way humans do.** When you asked about shaking hands, I provided a general response based on my training data, which includes a vast range of text from the internet and other sources. My responses are generated by analyzing the text of your question and providing a relevant answer based on the patterns I've learned during training.

# The Gödel theorem applied to AI

- «**In any consistent system...there are formulae which cannot be proved-in-the-system, but which we can see to be true.**» (Lucas 1961)
- Applies to programs/machines as well
- It follows that no machine can be a complete or adequate model of the mind, that minds are essentially different from machines.



John Lucas

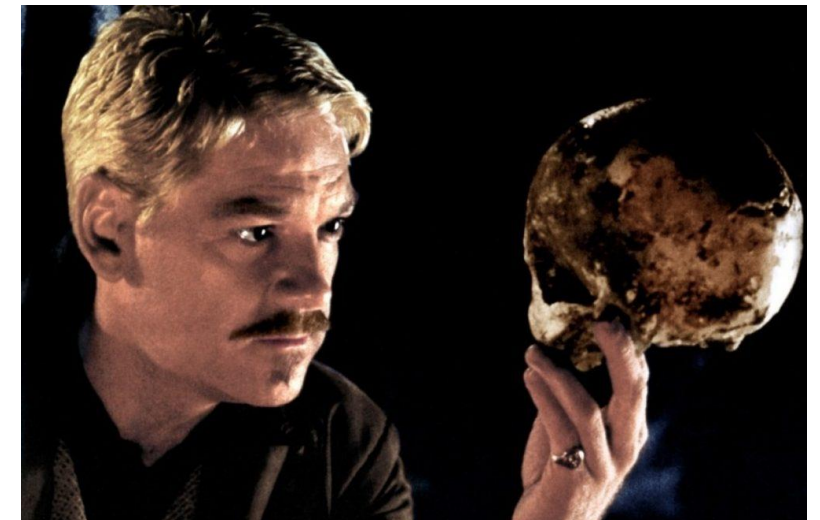
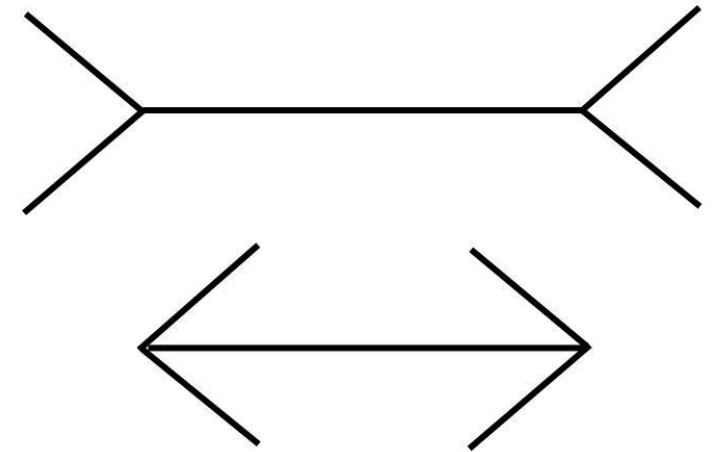
# Replies to Gödel-style arguments

- « [It is possible that] the **human mind** (in the realm of pure mathematics) is equivalent to a **finite machine** that, however, is **unable to understand completely** its own functioning.”  
(Gödel, 1995, p. 309)
- Maybe **cognition is characterized by a different formal system C** for which, however, a human **cannot prove its own G sentence** (Megill, Melvin, & Beal, 2014).
- What if the AI is allowed to be **inconsistent**? Can we prove that **we** are consistent?

# The framing problem (Fodor 1987)

- Encapsulated cognitive faculties: take only a selection of information into account (irrational)
- Unencapsulated cognitive faculties: « When to stop thinking? » (cf. Hamlet)
- Framing problem: how to non-arbitrarily break off (« frame ») thinking
- We somehow manage it, but nobody can formalize how, which would be necessary for constructing AGI

Muller - Lyer Illusion



# III. Is the Human Mind a Computer?

No, it isn't

- Gödelian-style arguments: Lucas 1961/2003; Penrose 1995
- Searle 1980: Chinese room argument
- Ross 1992/Feser 2013: immateriality of thought

# The Chinese Room Argument

- Imagine a man who knows only English sitting in a room
- He receives questions in Chinese
- Via a handbook (in English) he constructs answers in Chinese he doesn't understand
- Since the setup simulates a computer, computers don't understand





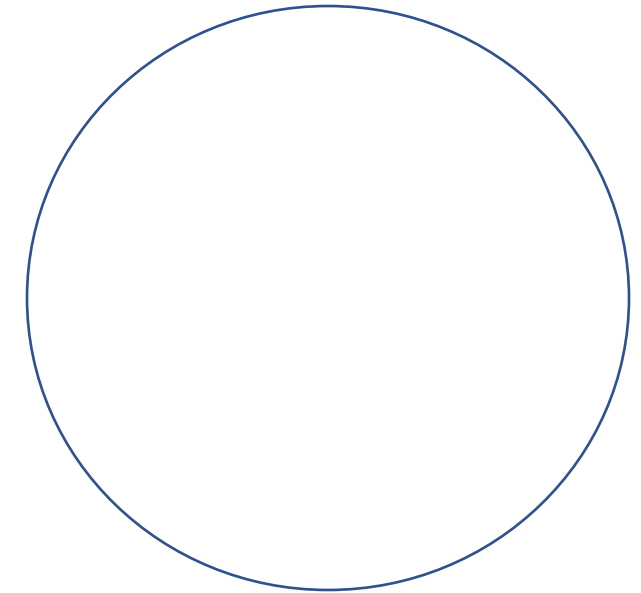
# What understands, if anything?

- Searle: understanding lies in **causal powers** of the brain
- Of course the man in the box doesn't understand Chinese, but the **system as a whole** does
- Parallel to physicalism-dualism debate: **redefining** mental terms  
(consciousness/intentionality)



# The immateriality of thought

- **P1: All formal thinking is determinate.**  
**P2: No physical process is determinate.**  
**C: Thus, no formal thinking is a physical process.**
- P1: In the sense that it determinately satisfies a certain function/conception.
- P2: Physical processes/objects are indeterminate between different impossible descriptions (e.g. circle could be a n-agon, mitosis could be schmitosis)



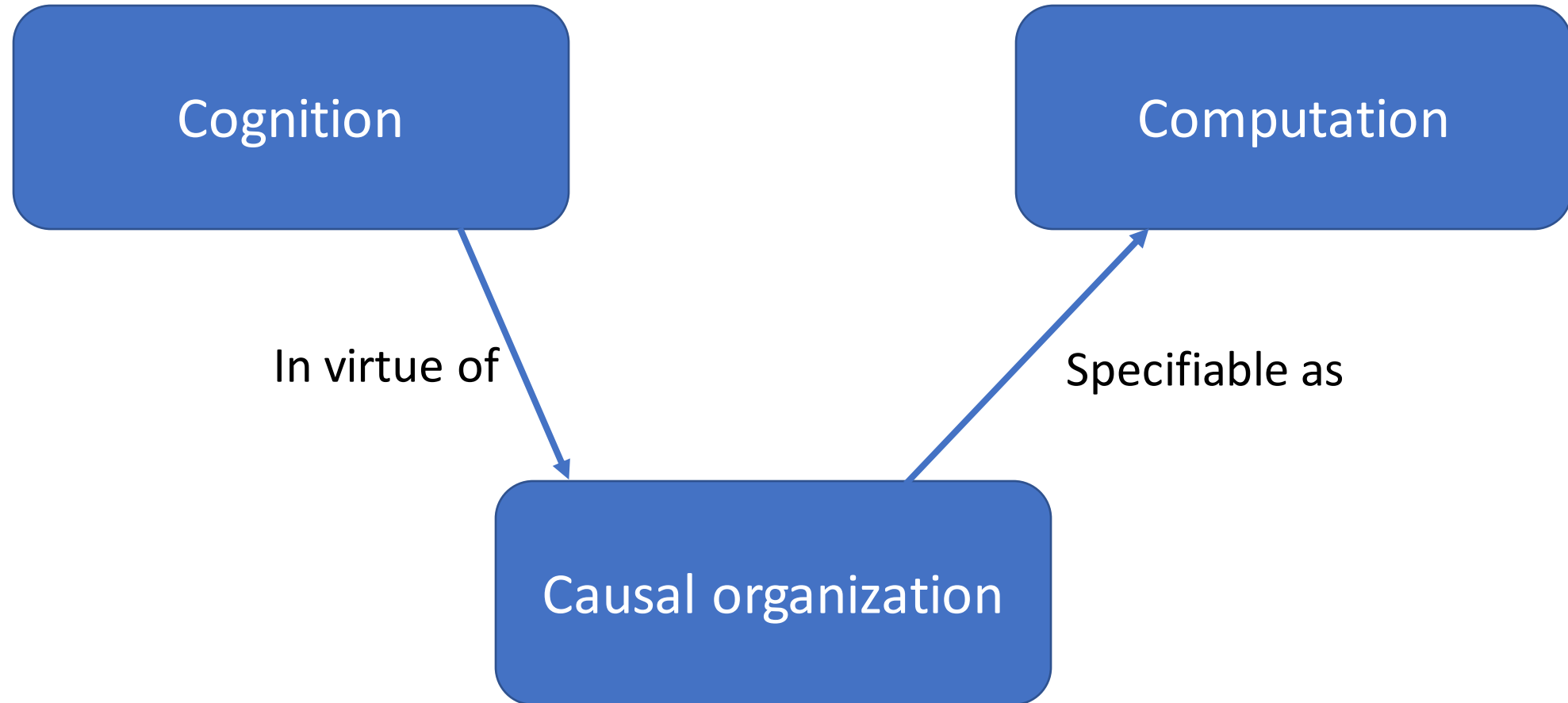
# The immateriality of thought

- **Thus, to the extent that computations depend on (are mirrored by) a physical substrate, the argument shows that the human mind isn't a computer**
- Further, are computations themselves determinate?  
(See plus/quus argument by Kripke)
  - $x \text{ quus } y = x + y$ , if  $x, y < 57$ ; = 5 otherwise.
  - How do we know if someone is following the plus or quus rule? We don't.
  - The day is saved if "plus" is a determinate concept **even distinct from people's mental events.**

# Yes, it is (Chalmers 2012)

- Thesis of **computational sufficiency**: right kind of computational structure suffices for mind
- Implementation of computational structure: “when the **causal** structure of the physical system **mirrors** the **formal** structure of the computation.» (isomorphism)
- Consequence: **every** (!) physical system implements *some* computation!
- Cognitive systems are cognitive ***in virtue of being computational***
- **Causal organization** (topology) is nexus btw. cognition and computation

Yes, it is (Chalmers 2012)



# Yes, it is (Chalmers 2012)

- Chalmers' argument rests heavily on his **distinction between psychological and phenomenal consciousness** and the acceptance of functionalism
- Further, that **link causal topology – computation** is „straightforward“
- Further, that **link cognition – causal organization** is „straightforward“
- Conspicuous: quickly dismisses Gödel-style arguments
- And does not even address Maudlin's (1989) worries point-blank directed at the triad of mind, causality and computation

# Yes, it is (Chalmers 2012)

- «there are some ways that empirical science might prove it to be false: (1) if the fundamental **laws** of physics are **noncomputable**... (2) if it turns out that our **cognitive** capacities depend essentially on **infinite precision** in certain analog quantities, ...(3) if... that cognition is mediated by some **non-physical substance** whose workings are not computable.»
- (1) could be denied if an Aristotelian metaphysics of nature (Koons 2022) is true (apart from that, what does it even mean for a law of physics to be computable?)
- (2) concepts *are* «infinitely precise». See Feser's argument
- (3) there are strong philosophical arguments as well as empirical evidence (NDEs) that cannot be swept under the rug



So, is it ?

Computational view of the mind requires refutation of

- Lucas/Penrose
- Dreyfus
- Maudlin
- Searle
- Feser/Ross
- Substance dualism