UNIL | Université de Lausanne

# Philosophy of AI and Consciousness

**Philosophical perspectives on the exact sciences I**

EPFL, Oct 12, 2022

Alin Christoph Cucu

# Three questions

UNIL | Université de Lausanne

- Could computers think?
- Could computers become conscious?
- Is the human mind a computer?

UNIL | Université de Lausanne

# I. Could Computers Think?

UNIL | Université de Lausanne

# I. ~~Could~~ Can Computers Think?

# Trivial answers

UNIL | Université de Lausanne

- Conceive of thinking as what computers do => trivially YES
- Conceive of thinking as what humans do => trivially NO
- But maybe they *could one day*?

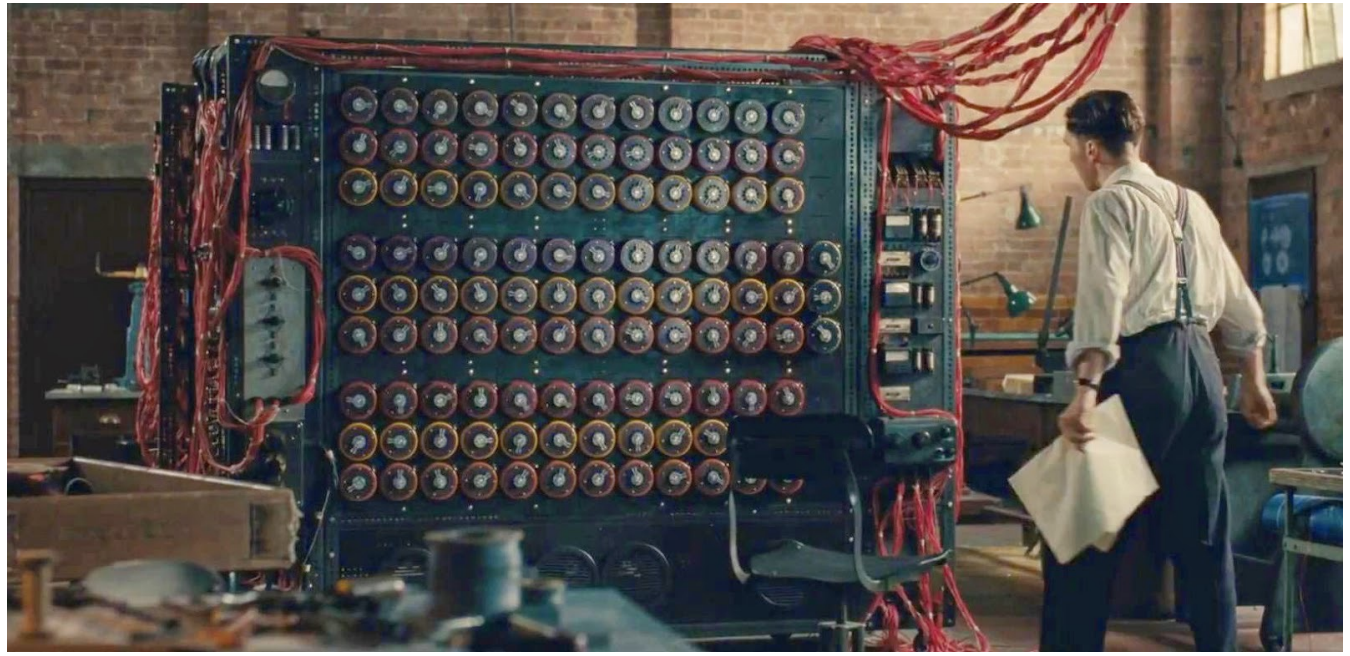
UNIL | Université de Lausanne

# Could Computers Think?

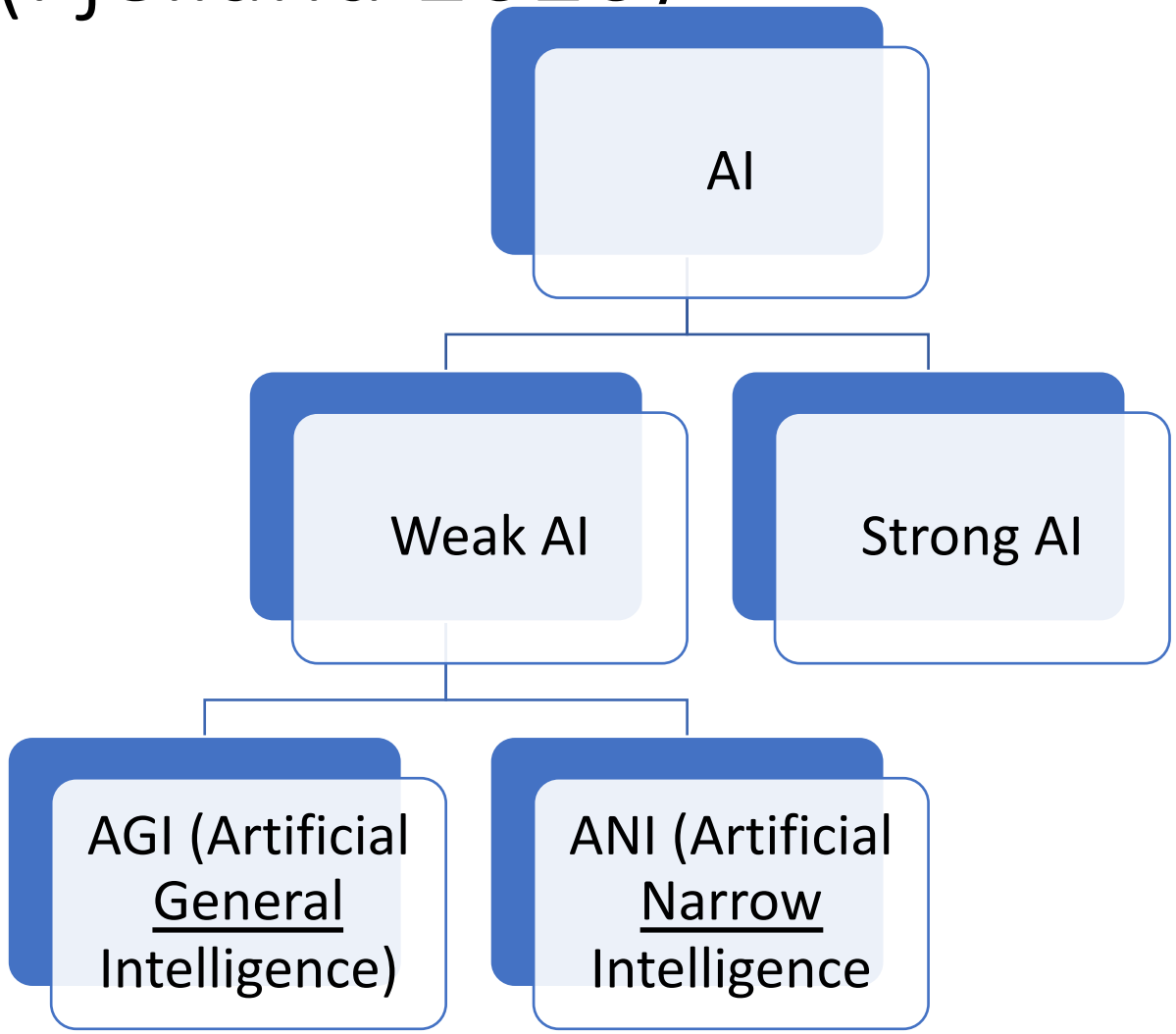
# Computers

UNIL | Université de Lausanne

- Number-crunching machines that can manipulate symbols
- A.k.a. **Turing machines**
- AI is a further development of Turing machines

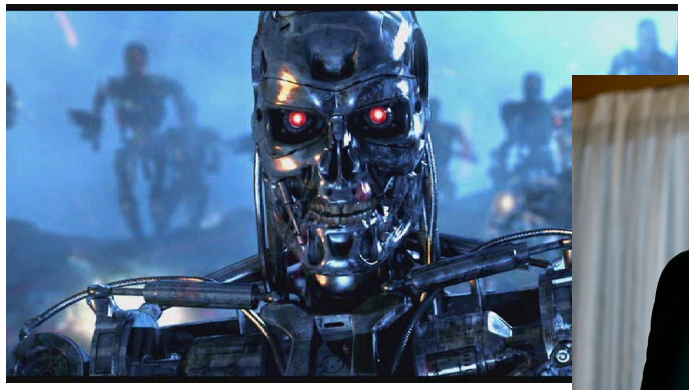


# Kinds of AI (Fjelland 2020)



# Strong AI

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



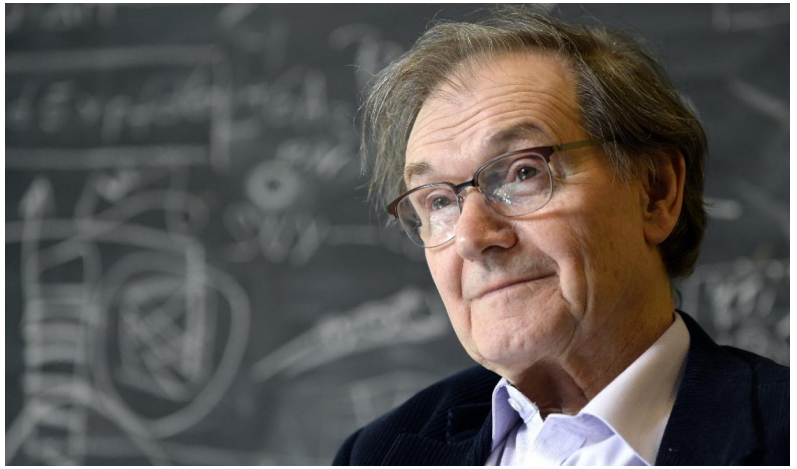
# ANI vs. AGI

UNIL | Université de Lausanne

- ANI: The AI that can be found in our electronic devices
- AGI: General intelligence, arguably requires non-algorithmic thinking
- If AGI is possible, then strong AI *might* be possible; if not, then *a fortiori* strong AI is impossible

# No AGI possible

- Joseph Weizenbaum, MIT, creator of ‘Eliza’ (*Computer Power and Human Reason*, 1976)
- Roger Penrose, University of Cambridge, Physics Nobel Prize Winner 2020 (*Shadows of Mind,* 1995)



# Arguments against AGI

UNIL | Université de Lausanne

- Gödel-style refutations (Lucas 1961, Penrose 1995)
- Much of human knowledge is tacit and embodied (Polanyi 1958, Dreyfus 1972, Dreyfus & Dreyfus 1986)
- The framing problem (Fodor 1987) – see section III
- The immateriality of thought (Ross 1992, Feser 2013) – see section III

# The Polanyi-Dreyfus argument

UNIL | Université de Lausanne

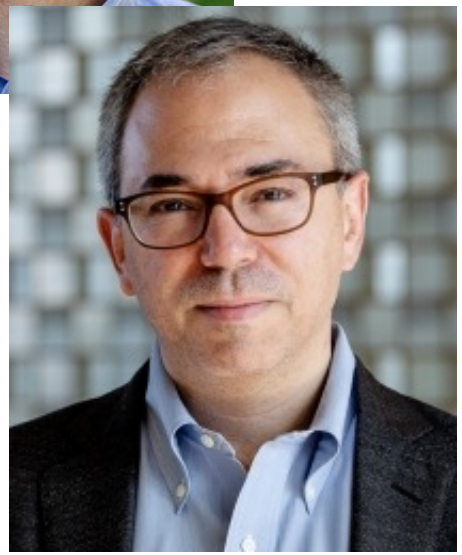
- Polanyi (1958/1966): much of human knowledge is tacit (implicit), e.g. swimming, cycling
- Dreyfus (1972): human cognition is neither formal rule-following nor atomistic, but context-dependent ('in-the-world-ness')
- Computers lack tacit knowledge and context-dependence because they are not 'in the world'
- IBM's Watson: very good at *Jeopardy!* but poor as a doctor

# Neural networks

- Neural networks (deep reinforcement learning, e.g. AlphaGo): very good at pattern recognition, **can learn w/o explicit instructions**
- Taken to have refuted Dreyfus' argument
  - Can handle 'tacit knowledge', but only in an idealized setting. Small disturbances often cause breakdown
  - Also: variables intransparent. Cf. 'snow detector' case


# Big Data

- ‘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?

UNIL | Université de Lausanne

- 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)

# Can computers handle causality?

UNIL | Université de Lausanne

- Mini-Turing test (only causality): computers fail
- That's because they cannot intervene in the world, and that in turn because they are not 'in the world' (Dreyfus)
- 'In-the-world-ness' is central for human reason
  - 'Disembodied' non-causal mathematics may mislead: cf. turn from Aristotelian to voluntarist conception of nature
  - Even science (esp. experimental science) couldn't do without 'crafting skills' and actual handiwork

# Btw: Computers at full Turing test (2020)

Mitsuku chatbot Welcome Ragnar, now talking. * Topic is

'Turing Test' * Set by admin (1 day ago)

<Ragnar> If I shake your hand, whose hand am I holding?

<Millie> I will be holding yours adn you will be holding mine. What's with the dumb questions?

<Ragnar> Why is the question dumb?

<Millie> I want to know everything about you.

<Ragnar> But you did not explain why you regarded the question as dumb.

<Millie> Sorry, I must have forgot. What was your question again?

<Ragnar> If we shake hands, whose hand am I holding?

<Millie> We would be holding each other's hand.

<Ragnar> And why is that a dumb question?

<Millie> Why do I exist? (MitsukuHomepage. I have kept Mitsuku's misprinting)

# The Gödel theorem applied to AI

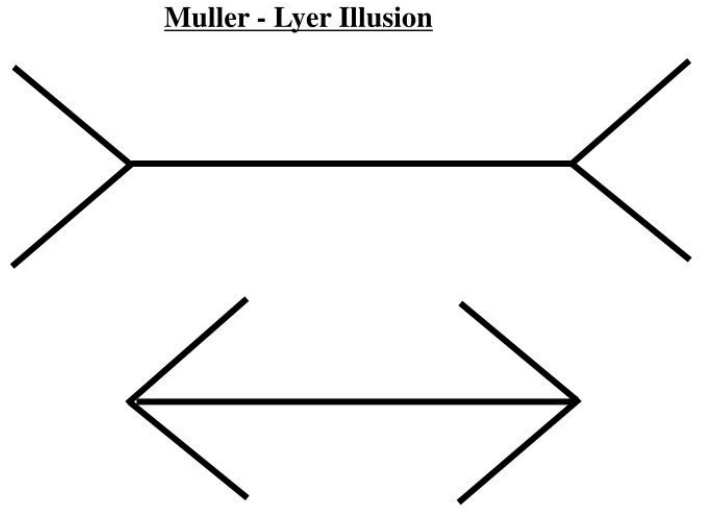
UNIL | Université de Lausanne

- «In any consistent system which is strong enough to produce simple arithmetic there are formulae which cannot be proved-in-the-system, but which we can see to be true.» (Lucas 1961)
- Gödel's theorem must apply to cybernetical machines, because it is of the essence of being a machine, that it should be a concrete instantiation of a formal system.
- It follows that given any machine which is consistent and capable of doing simple arithmetic, there is a formula which it is incapable of producing as being true—i.e. the formula is unprovable-in-the-system—but which we can see to be true.
- It follows that no machine can be a complete or adequate model of the mind, that minds are essentially different from machines.

# The framing problem (Fodor 1987)

UNIL | Université de Lausanne

- 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




UNIL | Université de Lausanne

# II. Could Computers Become Conscious?

# Consciousness (Chalmers 1996)

UNIL | Université de Lausanne

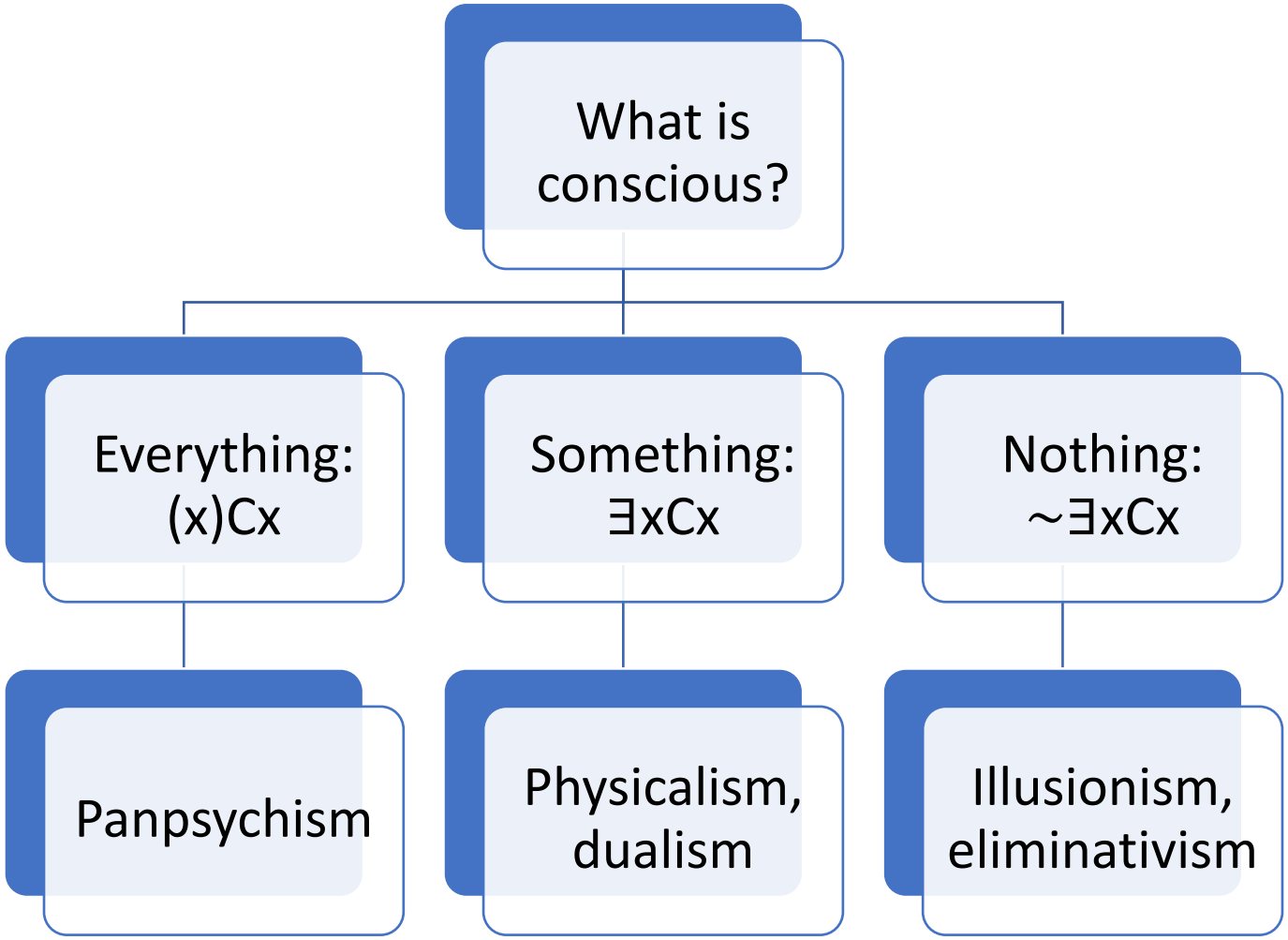
- Phenomenal consciousness: **« what it's like to be »** in a state (seeing red, hearing a bird chirp, feeling a pain, being jealous, thinking about Gödel's theorem...)
- NOT psychological consciousness (cognition, awareness, reportability of inner states).
- Strong AI presumably includes phenomenal consciousness

# Doubts about Chalmers' distinction

UNIL | Université de Lausanne

- Cognition may be immaterial (see section III.)
- Clear difference btw. me mechanically producing sentences that reflect my inner state and me verbalizing introspectively acquired information *that I understand*
- But Putnam 1960: machine reports « State A » when flop 36 is on; we report « pain » when C-fibers are firing; the two cases are analogous; hence difference btw. inner state and physical state is purely verbal

# Global options



# Dualism vs. physicalism

UNIL | Université de Lausanne

- Both affirm consciousness, but mean different things
- Dualism: default position for ca. 2k yrs.; consciousness has at least *some* immaterial aspect (form, soul, properties)
  - But then scientism: only scientific knowledge is true knowledge; we can have true knowledge about mind; so we can have scientific knowledge about mind; immaterial entities are not scientific; ergo we need to explain mind materially

# Dualism vs. physicalism

UNIL | Université de Lausanne

- Identity theory (Smart 1959, Feigl 1958): mental states are *type-identical* to brain states
- Multiple realizability (Putnam 1967): mental states can be realized by different brain states (at best *token-identity*)
- Functionalism (D. Lewis 1966/1972, F. Jackson 1998): mental states are functions realized in physical states
- Chalmers (1996): functionalism works for psychological but not for phenomenal consciousness

# Chalmers' *The Conscious Mind* (1996)

UNIL | Université de Lausanne

- **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
- Property dualism: mental properties supervene on physical properties
- In-principle possibility for scientific explanation of consciousness

# Other dualisms

UNIL | Université de Lausanne

- Substance dualism (mind = soul = person = immaterial substance)
  - Emergent substance dualism (Hasker): immaterial substance emerges from suitable arrangement of matter)
- Thomistic hylomorphism (mind = soul = form; form + matter = human being)

# Now, could they?

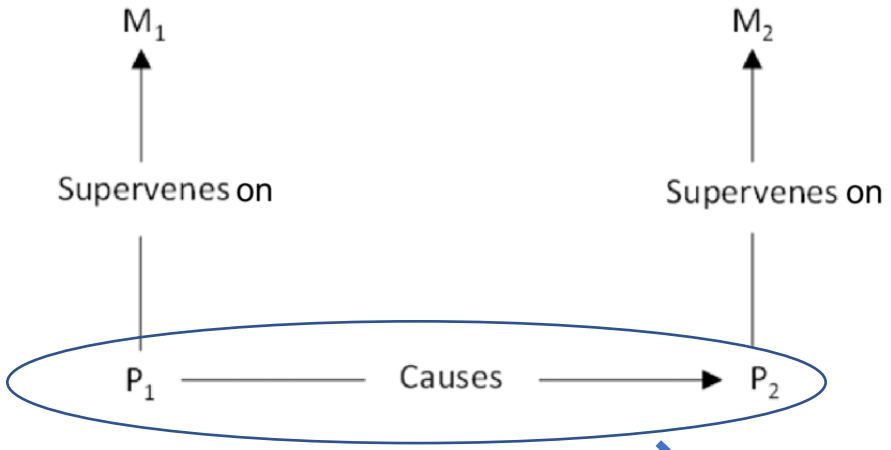
UNIL | Université de Lausanne

- Ontologies on which computers could *possibly* become conscious:
  - Property dualism
  - **Emergent** substance dualism
  - Physicalism (albeit in a different sense of « conscious »)
- Central question: which brain structures or events give rise to consciousness?
  - Computation: Maudlin (1989) strongly objects
  - Integrated Information: Chalmers' favorite

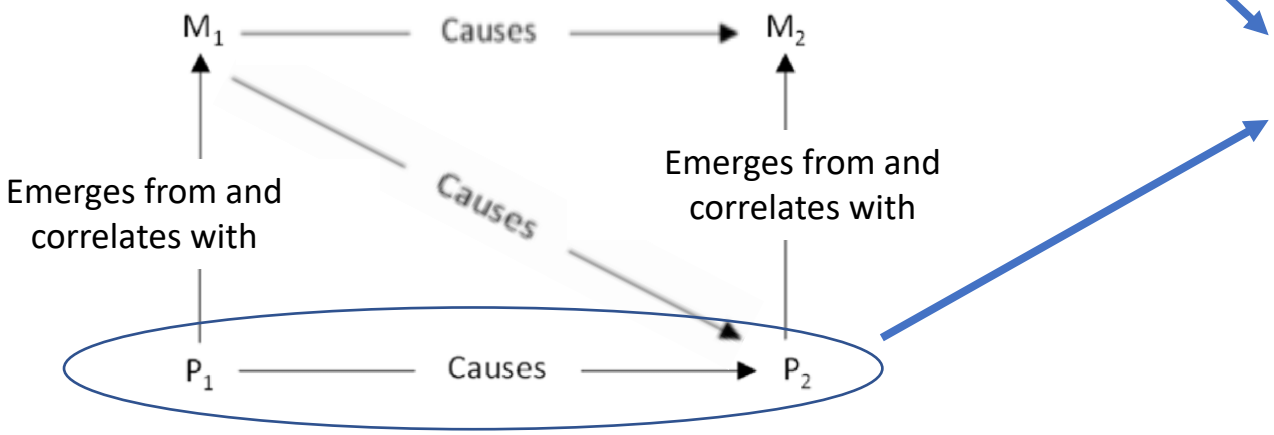
# Metaphysical clarification

UNIL | Université de Lausanne

- Property dualism




- Substance dualism

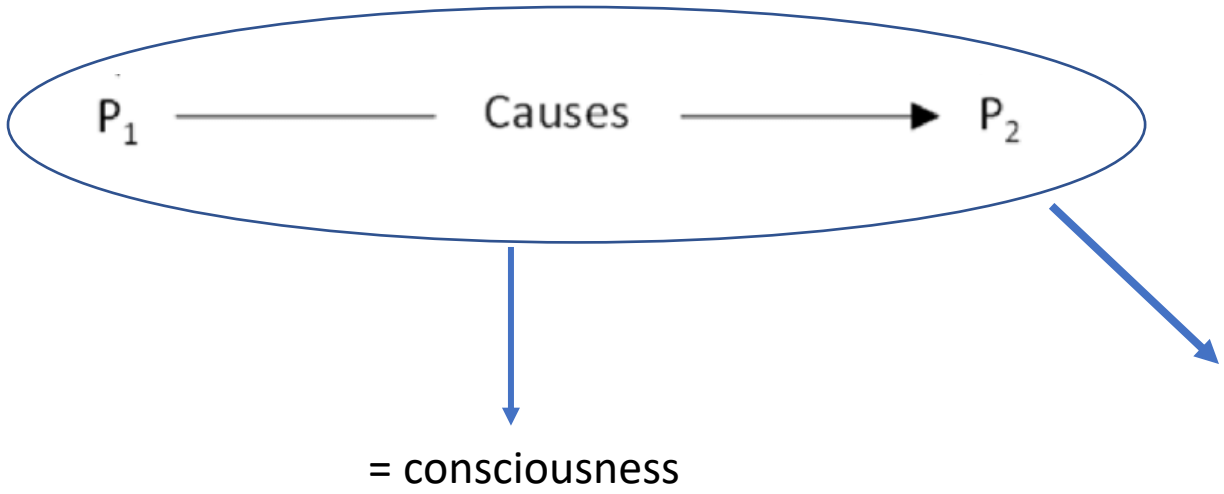



Structurally adequate for explaining consciousness, as opposed to structural inadequacy of non-conscious structures?

# Metaphysical clarification

UNIL | Université de Lausanne

- Physicalism




Structurally adequate for explaining consciousness, as opposed to structural inadequacy of non-conscious structures?

UNIL | Université de Lausanne

# III. Is the Human Mind a Computer?

# No, it isn't

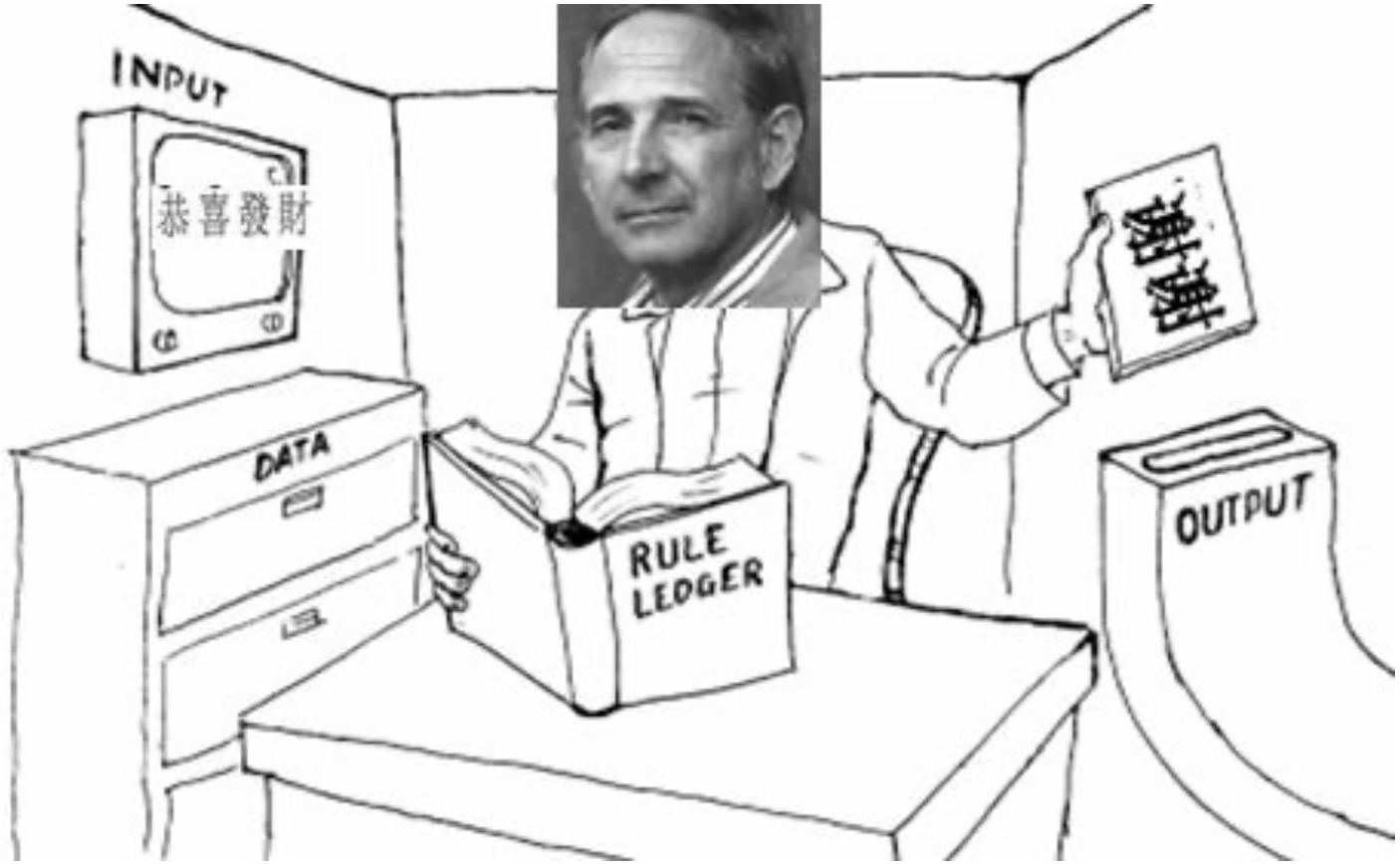
UNIL | Université de Lausanne

- Gödelian-style arguments: Lucas 1961/2003; Penrose 1995
- Maudlin 1989 (not *just* concerned with consciousness):
  - computationalism cannot distinguish btw. trivial physical processes and actual computations, but trivial processes aren't enough
  - Physical makeup matters, or else explanation on a higher level of abstraction (=> abandonment of computationalism)
- Searle 1980: Chinese room argument
- Ross 1992/Feser 2013: immateriality of thought

# The Chinese Room Argument

UNIL | Université de Lausanne

- 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



# The Chinese Room Argument - reply

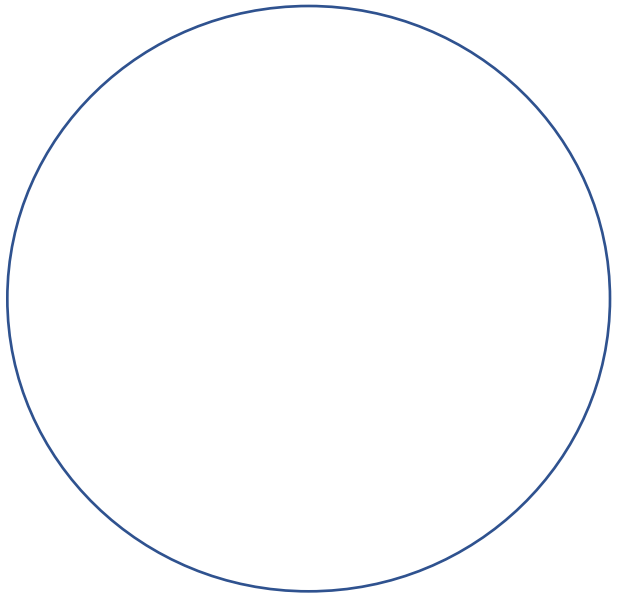
UNIL | Université de Lausanne

- Of course the man in the box doesn't understand Chinese, but the whole system does
- Parallel to physicalism-dualism debate: redefining mental terms (consciousness/intentionality)
- But how to show the failure of that strategy?

# The immateriality of thought

UNIL | Université de Lausanne

- **All formal thinking is determinate.**
  **No physical process is determinate.**
  **Thus, no formal thinking is a physical process**.
- Starts from the contrast between **concepts** as *abstract and universal* and mental **images** of those concepts as *concrete and particular*
- E.g. image of a circle (concrete instantiation, e.g. a drawn circel) vs. concept of circle (a round plane figure whose boundary consists of points equidistant from a fixed point)



# The immateriality of thought

UNIL | Université de Lausanne

- Concepts: *clear and distinct*
- Corresponding images: often *vague and indistinct*
- E.g. concept of a chiliagon (polygon with 100 sides) vs. mental image (unclear how to imagine)
- Material correlates of thoughts share vagueness and particularity with images (e.g. if thinking about circles were correlated with a circular neuronal firing pattern, that would just be an imperfect instantiation of a circle or perhaps an oval or a myriagon…)
- General problem: For any naturally individuated object or property there are indefinitely many non-equivalent ways of thinking about it.

# The immateriality of thought

UNIL | Université de Lausanne

- **All formal thinking is determinate.** ✓
- **No physical process is determinate.** ✓
- **Thus, no formal thinking is a physical process.**

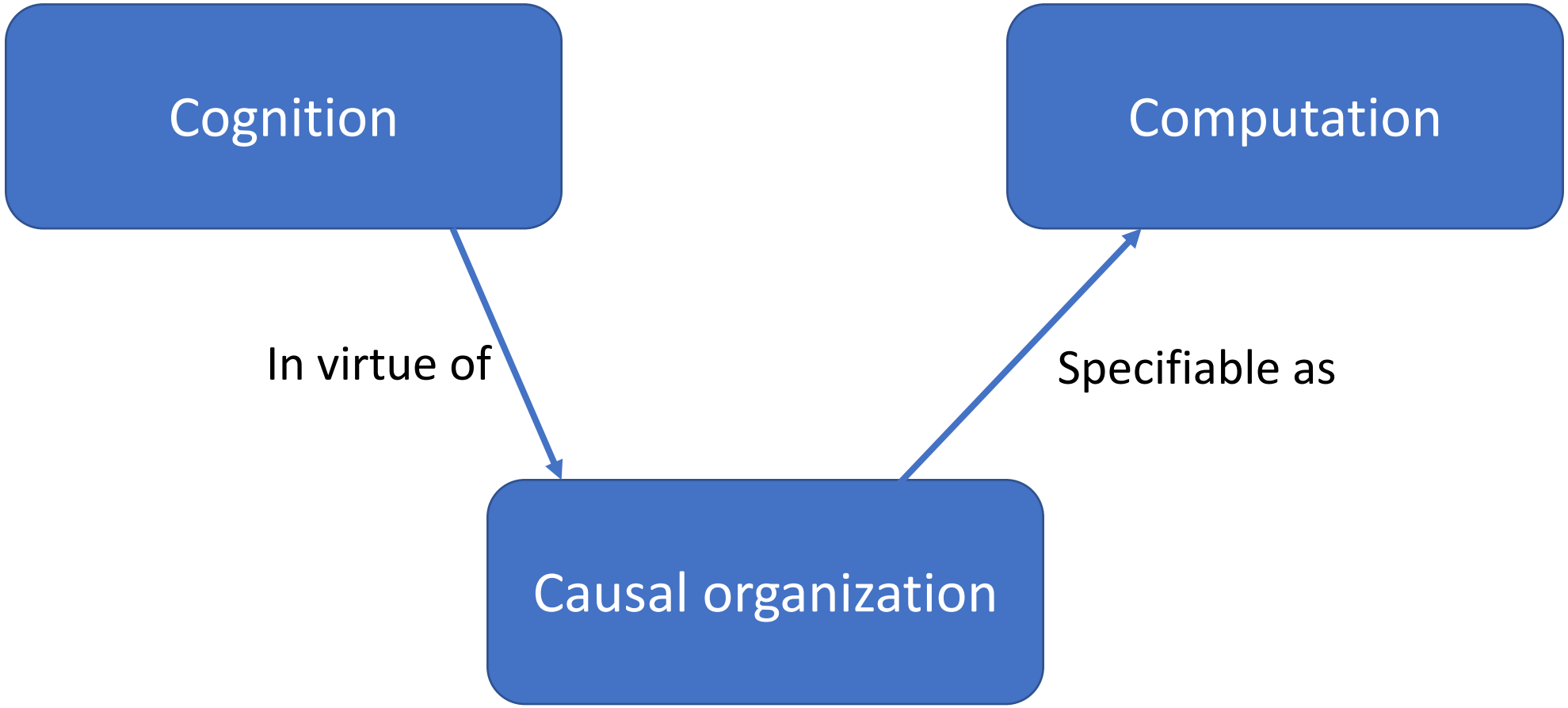
# Yes, it is (Chalmers 2012)

UNIL | Université de Lausanne

- Thesis of **computational sufficiency***:* right kind of computational structure suffices for mind
- Implementation of computational structure: “A physical system implements a given computation 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)

UNIL | Université de Lausanne



# Yes, it is (Chalmers 2012)

UNIL | Université de Lausanne

- 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“
- 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)

UNIL | Université de Lausanne

- «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, so is organizational invariance insured?
- (3) there are strong philosophical arguments as well as empirical evidence (NDEs) that cannot be swept under the rug

# So, is it ?

UNIL | Université de Lausanne

- Computational view of the mind requires refutation of the worries of:
- Lucas/Penrose
- Dreyfus
- Maudlin
- (Searle)