Philosophy of Mind

Philipp Koehn

28 January 2025

mind

René Descartes

- French philosopher, mathematician, and scientist, 1596–1650
- One claim to fame:
  Cartesian coordinate system
- 1637: Discourse on the Method
- 1641: Meditations on First Philosophy
- Historical context: In 1633 Galileo was condemned by the Catholic Church

First Meditation

LAYING OUT THE PLAN

- *I must once for all seriously undertake to rid myself of all the opinions which I had formerly accepted, and commence to build anew from the foundation*

- *It is not necessary that I should show that all of these are false ... I ought no less carefully to withhold my assent from matters which are not entirely certain.*

- *All that up to the present time I have accepted as most true and certain I have learned either from the senses or through the senses; but it is sometimes proved to me that these senses are deceptive, and it is wiser not to trust entirely to anything by which we have once been deceived.*

First Meditation

QUESTIONING EVERYTHING

- Some things seem certain. *For example, there is the fact that I am here, seated by the fire, attired in a dressing gown, having this paper in my hands and other similar matters. And how could I deny that these hands and this body are mine*

- But: I have dreamed similar things. *There are no certain indications by which we may clearly distinguish wakefulness from sleep.*

- *I shall then suppose, ... some evil genius ... has employed his whole energies in deceiving me; I shall consider that the heavens, the earth, colours, figures, sound, and all other external things are nought but the illusions and dreams of which this genius has availed himself in order to lay traps for my credulity.*

- *I may at least do what is in my power [i.e. suspend my judgment], and with firm purpose avoid giving credence to any false thing, or being imposed upon by this arch deceiver*

Second Meditation

WHAT REMAINS: I EXIST

- *I suppose, then, that all the things that I see are false; I persuade myself that nothing has ever existed of all that my fallacious memory represents to me. I consider that I possess no senses; I imagine that body, figure, extension, movement and place are but the fictions of my mind. What, then, can be esteemed as true? Perhaps nothing at all, unless that there is nothing in the world that is certain.*

- *Was I not then likewise persuaded that I did not exist? Not at all; of a surety I myself did exist since I persuaded myself of something*

- *We must come to the definite conclusion that this proposition: I am, I exist, is necessarily true each time that I pronounce it, or that I mentally conceive it.*

- *I am, however, a real thing and really exist; but what thing? I have answered: a thing which thinks.*

I think, therefore I am.

(René Descartes, Discourse on the Method, 1637)

Properties of the Mind

- Qualia: sensations or sensory qualities (e.g., how the color red looks like)
- Intentional: having agency, beliefs, goals
- Feelings, emotions
- Personality, character
- Consciousness
- Subjective (cannot be observed from the outside)

Other Minds

- To function in society, we constantly infer mental states of others
  - we note pleasure or annoyance to guide our interaction
  - but we are sometimes mistaken about other's beliefs

- Arguments for other minds:
  - it is intuitive and common sense
  - once I associate my mind with my body,
    I pretty much assume other minds for other bodies
    (we constantly define what we are in contrast to others)
  - analogy:
    i have a body, others have a body,
    i have a mind → others have a mind
  - it is the *best explanation* (simplest)

Minds of Animals?

- Who has a mind?
  - your dog or cat (not a human)
  - a pig (not a pet)
  - a fish (not a mammal)
  - an ant (not a vertebrate)
  - a worm (no nervous system)
  - a tree (not an animal)
  - grass (no annual life cycle)
  - a bacterium (not a multi-cell organism)
  - a virus (not a single-cell organism)
  - water in the ocean (not animate)
  - fire (not a physical object)
- Where do we draw the line?
- Possible requirements: language, tool use, memory, nervous system, ...
- Is there a line or a gradual transition?

Minds of Computers?

- If a computer behaves like a human, should we infer a mind?

- The main argument against a mind in computers:
  - the functioning of a computer follows strict rules
  - we can fully understand and explain these instructions
    (in theory, in practice it will be hard to track down exact causes of actions)
  - there is no need for a ”mind” for it to function

- However:
  - the functioning of our body follows physical laws
  - maybe one day, we can fully account for all the biochemical processes
  - but we have a mind

- More detail later...

Free Will

- Free will: the ability to choose a course of action
- Basis of morality
  - if you had no choice, you cannot be judged
  - inflicting pain and punishment for misdeeds is justified
  - criminal justice system has notions such as ”intent”, ”mentally capable”
- Opposed to notions of destiny or fate (or ”determinism”)
- Do we have free will?
  - intuitively, yes
  - but there is no proof
    (we never know if we could have really decided otherwise)

matter

The Material World

- Do we live in a computer simulation?

  Are you just imagining this?

  Maybe, but what difference does it make?

- Science = inquiry into properties of the world

- Scientific method
  - direct experience through our senses or use of instruments
  - hypothesis that makes predictions
  - validation of predictions in experiments

- Scientific theory
  - the universe is very large
  - processes in the world follow few, strict rules

Science

- Properties of the material world
  - 4-dimensional spacetime
  - matter (protons, neutrons, electrons, photons, ...)
  - energy and force

- Physical laws
  - conservation of mass, energy, etc.
  - laws of motion
  - gravity and relativity
  - thermodynamics
  - electromagnetism
  - photonics
  - quantum mechanics
  - radioactivity

- Pretty solid understanding, but not a comprehensive universal theory (yet)

- Application of science (technology) hard because of complexity of real world

Known Limits of Science

- Randomness
  - radioactive decay
  - quantum states
- Limits to knowledge
  - Heisenberg's uncertainty principle: exact position and momentum of particle cannot be known
  - Observer effect: measuring something may change it
- Limits to logical systems
  - Goedel's incompleteness theorems
  - In any logical system, there will be statements that cannot be proven
  - real world example: you cannot write a perfect debugger
- But: we have no evidence for super-natural effects

the mind-body problem

Mind as Substance

- Descartes views the mind as a ”substance”
- Mind is a real thing (the ”soul”)
- Distinct from material substance
- Exists independent of the body
- A very common sensical view

Afterlife

- What happens when our body dies?

- It is hard to grasp the idea that our mind will disappear

- Christianity
  - if you lived a life without sin (or repented your sins), your soul go to heaven
  - if you lived a life of sin, your soul goes to hell
  - limbo = someplace inbetween (for instance, for infants)
  - purgatory = a place where the soul is cleansed before entering heaven

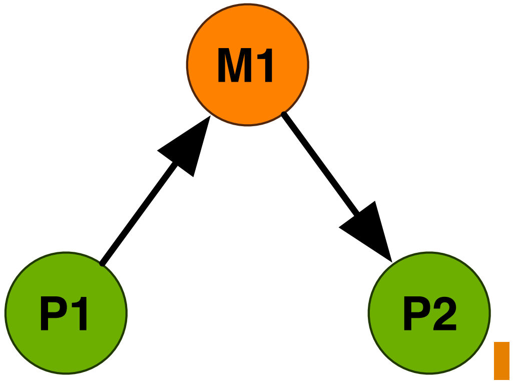
- Reincarnation
  - after death, your soul becomes attached to a newborn (maybe even an animal)
  - belief in Hinduism and other religions

Substance Dualism

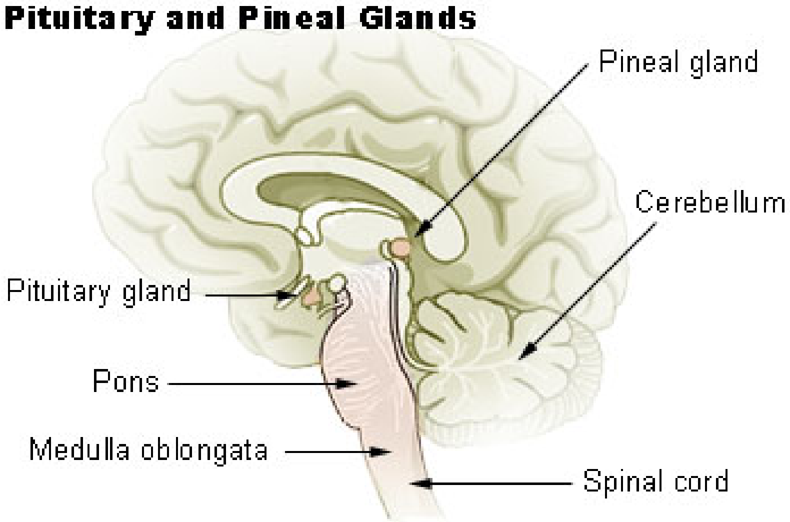
- Mind and body can exist independently
- Human = mind + body
- Ghost = just mind
- Zombie = just body

The Problem with Dualism

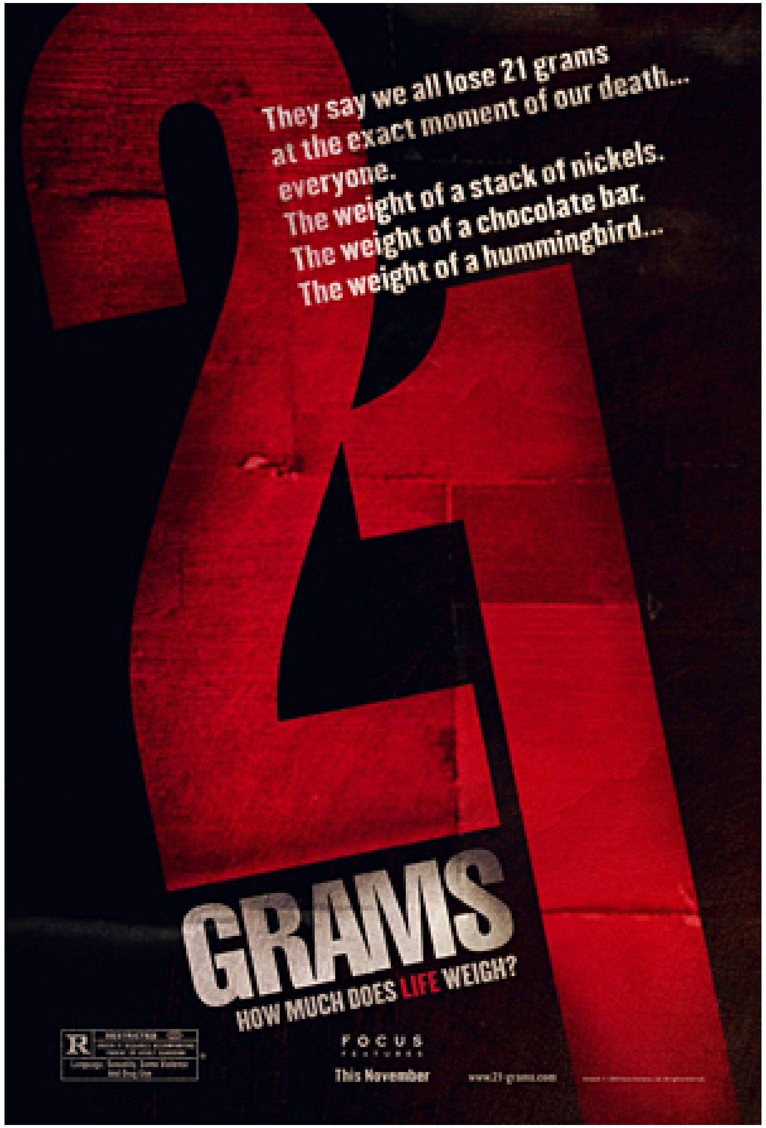
- The mind exists
- The physical word exists
- How do they interact?



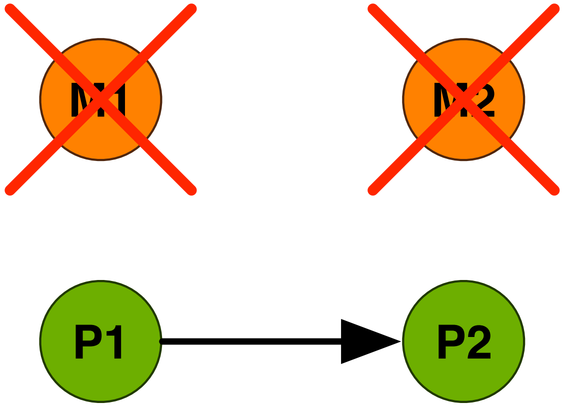
- However: if there are mental effects on physical states
  - they break current physical laws
  - they should be able to be measured
  - ... but none have been found so far (maybe in the pineal gland?)
  - ”ghost in the machine”



Mind as Matter

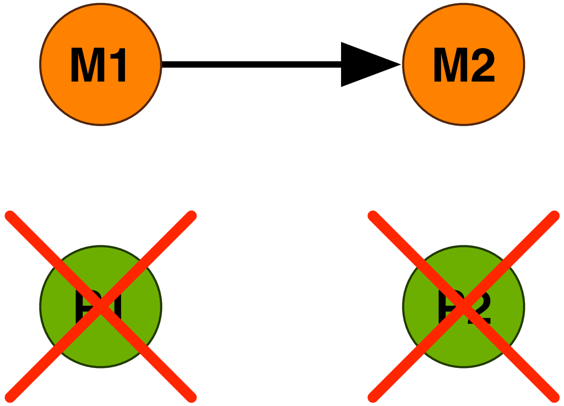

21
They say we all lose 21 grams
at the exact moment of our death...
everyone.
The weight of a stack of nickels.
The weight of a chocolate bar.
The weight of a hummingbird...
GRAMS
HOW MUCH DOES LIFE WEIGH?
R
FOCUS FEATURES
This November
www.21-grams.com

Materialism (or Physicalism)



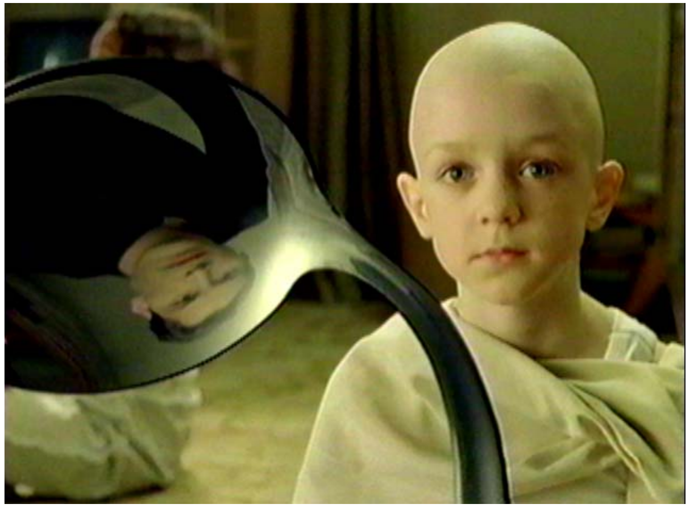
- There are no mental states that are not also physical states
- Every explanation of the world is grounded in physics
- You do not exist, it's just a bunch of biochemical processes

Idealism

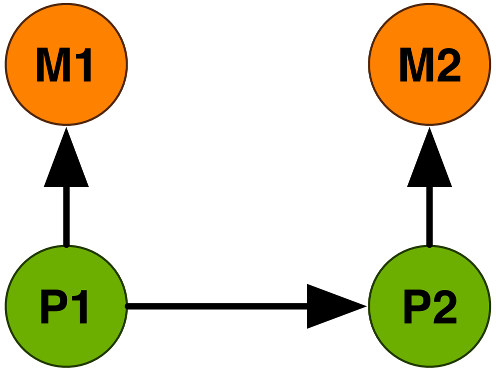


- Only the mind exists, the physical world only exists in our imagination (e.g., *"There is no spoon"*, The Matrix)

- Close to solipsm, the philosophical idea that only your mind exist



Epiphenomenalism



- Mental states completely depend on physical states
- But they have different properties (property dualism)
- We have consciousness, but conscious control is an illusion

Functionalism

- Human behavior can be explained by
  - the physical environment
  - mental states
- Brain replacement experiment
  - replace, one by one, each neuron with an electronic functional equivalent
  - at the end, we have an electronic simulation of the brain
  - at what point will the brain's owner report diminished consciousness?
- Functional equivalent = mental equivalent
- Dispute, if this implies epiphenomenalism

turing test

Can Machines Think?

- ”Imitation Game” proposed by Alan Turing (1950)

- Setup
  - a human (A), being honest about being a human
  - a computer (B), pretending to be a human
  - an interrogator (C)

- Performing the test
  - C is in a different room, communicates by typewriter
  - C has to find out if A or B is the human
  - C may ask any kind of question

- If C cannot tell the difference, then B is deemed to be able to ”think”

Turing Test Today: Loebner Prize

- Annual competition since 1990

- Restrictions
  - time limit: initially 5 minutes, now 25 minutes
  - restricted domain
    e.g., romantic relationships, Shakespeare’s plays, Burgundy wines

- Occasional press reports of passing this test

Passing the Turing Test

- Key:
  - make rambling, whimsical statements
  - drive the conversation by asking questions
  - have generic default responses

- Not very useful to push AI forward...

  (Marvin Minsky called the Loebner prize ”obnoxious and stupid”)

- Not clear if modern large language models can pass the test
- They may be too collaborative and helpful

Turing Test and Having a Mind

- Turing’s Thesis:

  if input/output behavior is equivalent,
  we have to attribute similar status of ”having a mind”

- Computationalism (”Strong AI”)
  - mind = information processor
  - inference in a model of cognitive processes = cognitive processes

Chinese Room

- Assume that you have a ”story understanding” machine
  - machine is given the text of a story
  - you can ask questions about the story, the machine answers them

- Example
  - *A man goes into a restaurant and orders a hamburger. When it arrives, he is very pleased, and when he leaves, he leaves a big tip for the waiter.*
  - Question: *Did the man eat the hamburger?*
  - Answer: *Yes.*

- Similar to the Turing Test, a machine passing this test can ”think”

Chinese Room Deconstructed

- Assume that machine operates in Chinese

- The machine executes a number of processing steps on the given input

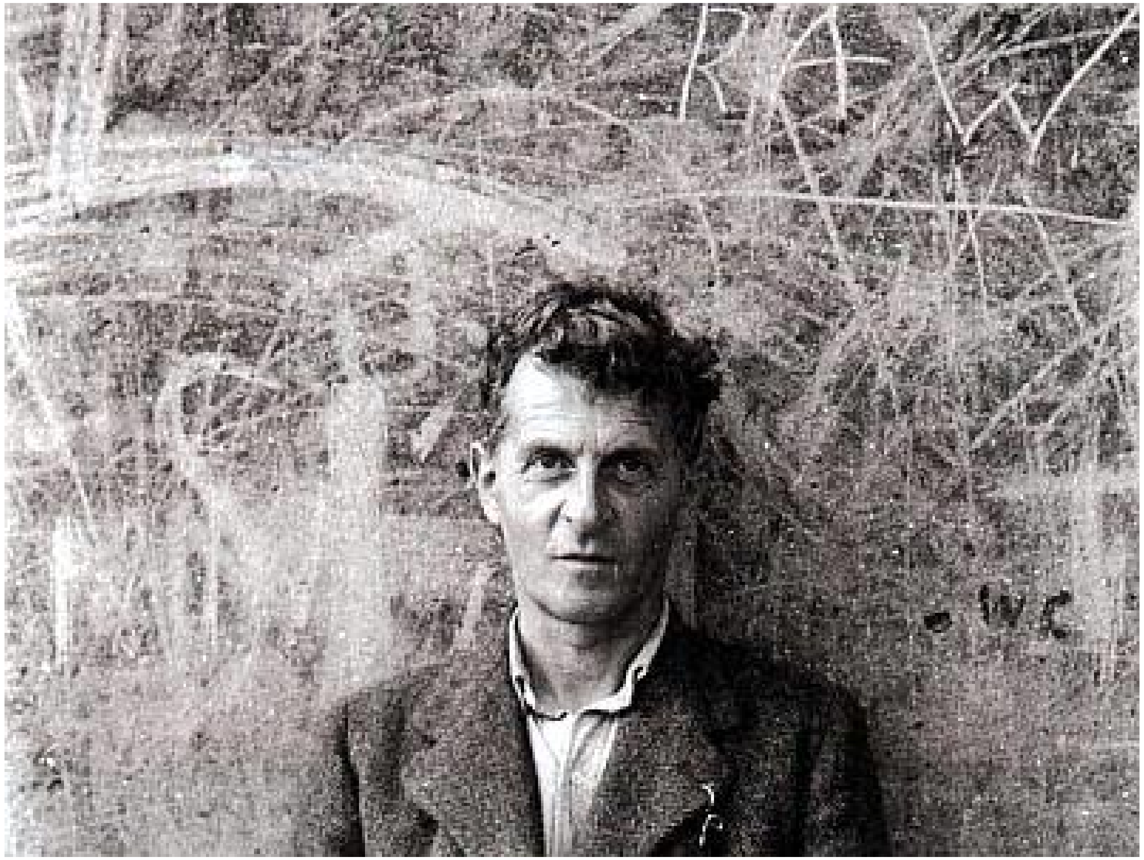
- This can be simulated by a ”Chinese room” which contains
  - a rule book
  - cards with Chinese symbols for the text, questions, and answers
  - a person (who does not know Chinese) following the rule book
  - a notebook to store intermediate processing results

- Intuition
  - the person operating the rule book, has no understanding of the story

  ⇒ there is no understanding — in either the Chinese room or the machine

wittgenstein

Ludwig Wittgenstein



- Born 1889
- Solved philosophy 1918 (Tractatus Logico-Philosophicus)
- Worked as school teacher
- Acccepted position at Cambridge in 1929, realized he had not solved philosphy (Philosopical Investigations, unfinished, posthumously published in 1953)

Tractatus Logico-Philosophicus

1. The world is everything that is the case.

2. What is the case (a fact) is the existence of states of affairs.

3. A logical picture of facts is a thought.

4. A thought is a proposition with a sense.

5. A proposition is a truth-function of elementary propositions. (An elementary proposition is a truth-function of itself.)

6. The general form of a proposition is the general form of a truth function, which is: $[\bar{p}, \bar{\xi}, N(\bar{\xi})]$. This is the general form of a proposition.

7. Whereof one cannot speak, thereof one must be silent.

Implications

- All facts about the world can be grounded in formal logic
  → we will explore such formal systems in later lectures

- *Most of the propositions and questions to be found in philosophical works are not false but nonsensical.*
  → *meaning of life* is not defined, just as *what color are your thoughts?*

- *Whereof one cannot speak, thereof one must be silent.*
  - there is no point in abstract discussions that are not grounded in fact
  - this also means that the whole of the tractatus is non-sensical, which is addressed as

    My propositions are elucidatory in this way: he who understands me finally recognizes them as senseless, when he has climbed out through them, on them, over them. (He must so to speak throw away the ladder, after he has climbed up on it.) He must surmount these propositions; then he sees the world rightly.

Do Words Really Describe Objects?

- We cannot define words

- Wittgenstein's example: the word *game*
(actually the German *Spiel*, which also means *play*)
  - there are physical games (sports), and mental games (chess)
  - there are also child's games, which may not be competitive
  - there are games you play by yourself

  ⇒ pretty much any property of *game* is violated by one kind of game

- Still, it is a useful word

Meaning of Words

- Language does not describe facts, it is used to communicate
- *The meaning of a word is its use in the language.*

Philosophical Investigations, 43

- The various uses of words can be best understood as *family resemblance*
  - use A is similar to use B, because they share trait X
  - use B is similar to use C, because they share trait Y
- If a word is used in a new context, we draw on the various uses in other contexts

Language Games

- Language is always used as part of an activity

- Each of these activities is a *language game*

- Right now, we play the *"professor instructs students"* game

- In a builder game, one worker may say *hammer*
  - He is not describing an object
  - He is issuing the request:
    Please give me the hammer that I need now to accomplish this task.

- Language as a whole is the sum of these language games

Implications

- Our thoughts are build on language
- Philosophy will never produce a theory
- *Philosophy just puts everything before us, and neither explains nor deduces anything. Since everything lies open to view there is nothing to explain.*

Philosophical Investigations, 126

The AI Problem

- Discussions about *thinking/mind/consciousness* of machines
  - are useful to illuminate the problem (”*puts everything before us*”)
  - also give us some insights into the problem of modeling knowledge
  - but will not reach conclusive facts

- We may be just playing a lot of definitional games around these words

- For our problem of artificial intelligence, we are really asking:

  How should we treat robots that resemble humans?