Philosophy and AI
Lecture 5: Philosophy of Mind I

Marco Degano

18 February 2025

Readings

Required:

- Glymour 2015: chapter 12, pp. 287 – 298
  - Required: Introduction; Some Metaphysical Views; Personal Identity; Reduction; The Intensional and the Extensional. Functionalism, Physicalism and the Cartesian Fallacy.

Optional:

- *The Mind–Body Problem*, Jonathan Westphal, Chapters 1 – 3
- *Minds and Machines* MIT OCW, Part 2, From Dualism to Functionalism
  `https://openlearninglibrary.mit.edu/courses/course-v1:MITx+24.09x+3T2019/course/`

Outline

Outline

Introduction

- Philosophy of mind explores the nature of the mind and its relationship with the physical world.
- The central question is the mind-body problem: how does the mind (or the mental) relate to the body and the physical (or the non-mental)?

Introduction

The core questions are:

- Are our minds, including consciousness, emotions, and decisions, ultimately reducible to physical systems, like neural networks or computational models?
- Or is the mind something fundamentally distinct, irreducible to physical phenomena?
- If mind and body are the same, why does the mind seem so distinct from physical matter?
- If they are different, how can the mind influence the body (e.g., intention leading to action) and how do physical events (e.g., sensory input) influence mental states?

Introduction

Related questions include:

- Personal identity: What defines continuity in a person's identity over time? Could AI systems have "personal identity" if they store memories and adapt over time?
- The hard problem of consciousness: Why do cognitive processes like perception and planning come with a subjective experience? Could AI ever develop this?
- What are thoughts and mental states, and how are they represented? Is it possible to represent these in an artificial system?
- Do thoughts require language, or can they exist independently? This is crucial for designing AI systems capable of reasoning without natural language.

Today, we'll primarily focus on the mind-body problem but also touch on these interconnected questions.

The Mind-Body Problem

Scenario: The teleporter scans you, disassembles (destroys) your original, and then reassembles you at the destination.

- Is the person who appears at the destination truly you, or just an identical copy?

The Mind-Body Problem

- At what point does the 'you' that is being reassembled differ from the original 'you'?
- If two identical versions of you emerge, which one is the real you?
- The original is kept, which of the two should be considered you?
- Does the medium (biological vs. digital) in which you exist affect your identity?

Outline

Views on the mind-body problem

- Dualism: The mental and the physical are fundamentally distinct, existing as two separate realms.
- Monism: The mental and the physical are of the same kind. Two main forms:
  - Physical monism (Materialism/physicalism): The mental is reducible to physical processes, such as neural or computational activity.
  - Idealistic monism: The physical world is a construct of the mental, built from perceptions and ideas.

Example Dualism: Descartes

- Mental entities (thoughts, intentions) are fundamentally different from physical entities (objects, bodies).
- Mental and physical entities causally influence each other:
  - A mental intention (e.g., deciding to move your arm) causes physical movement.
  - A physical entity (e.g., a chair) causes a mental perception when you look at it.
- The mystery of dualism: How can two completely different kinds of substances interact? (This is a key criticism of Descartes' theory.)
- Implication: Minds could exist independently of physical bodies, and vice versa.

Relevance to AI:

- Can we think of a 'mind' of an AI system that is independent of its physical hardware?

Example Physical Monism: Hobbes

- Mental activities like reasoning and judging are forms of computation.
- These computations are implemented physically by the body.
  - Hobbes compared mental processes to the symbolic manipulation of an abacus, where small components represent external reality.
- The mind is just a physical entity, functioning in a specialized way.
- This view is a strong form of physical monism known as materialism or physicalism.

Relevance to AI:

- This view aligns closely with AI research: cognitive functions are treated as computations performed by artificial systems

Example Idealistic Monism: Kant

- Physical entities are syntheses of mental entities, constructed from basic mental concepts:
  - For example, trees are mental combinations of "green," "brown", "tall", and other sensory perceptions.
- When we talk about physical objects, we refer only to things synthesized from our sensory experiences (which are mental in nature).

Outline

Background: Substances in Metaphysics

- Before asking how mental entities (thoughts, mental states) relate to physical entities (stones, chairs, bodies), it is important to first ask what entities are.
- This question is central to metaphysics, the branch of philosophy that explores the fundamental nature of reality.
- A classic idea originates from Aristotle:
  - Consider a chair: What makes it the same chair even if its properties (color, weight, height) change?
  - Many properties are contingent, meaning they can change without altering the object's identity.
  - The substance of an object refers to what remains when all its contingent properties are removed-a kind of name tag for its identity.

Background: Substances in Metaphysics

Different philosophers proposed varying views on substances:

- Aristotle: One kind of substance underlies both mental and physical entities.
- Descartes: Two distinct substances-mental substances and physical substances.
- Locke: Only physical substances exist.
- Berkeley: Only mental substances exist.
- Hume: Neither mental nor physical substances exist; identity is a mental construct.

Though abstract, this debate has practical implications, especially for understanding personal identity.

Background: Personal Identity

- How can we determine that past-you is identical to present-you?
  - For example, how do we know a convict is the same person who committed the crime?
- If entities, including persons, have substances, this question has a well-defined answer:
  - The underlying substance (the "name tag") must remain the same across time.
  - Even if other properties change (e.g., appearance, memories), the substance ensures continuity of identity.

Background: Personal Identity

- Without a concept of substance, identity becomes less clear:
- Consider the Ship of Theseus thought experiment:
  - Starting with a wooden boat, replace each plank one by one while simultaneously reassembling the original planks into a second boat.
  - Which is the same ship as the original?
    - If physical continuity defines identity, the boat with the replaced planks is the original.
    - If constituent material defines identity, the reassembled boat is the original.

Relevance to AI

- Do AI systems possess an essential substance that defines their identity?
  - Hardware configuration?
  - Software architecture?
  - Training data continuity?
- Does downloading and running a model create a new entity or preserve identity?
- When do gradual updates fundamentally change identity? (analogy with Ship of Theseus)
- If an AI evolves or adapts over time, can it be held responsible for earlier decisions?

Outline

Argument against reduction/physicalism: Brentano

- Brentano proposes an argument challenging the idea that the mental can be reduced to the physical (i.e., physicalism).
- The structure of the argument is as follows:
  - (1) The language of physics is extensional
  - (2) The language of mind is intensional
  - (3) If the mental reduces to the physical, true claims in mind-language must be derived from true claims in physics-language plus definitions.
  - (4) No addition of definitions to an extensional language can create an intensional language.
  - (C) Hence: The mental cannot be reduced to the physical.

Extensional vs Intensional Language

What does it mean for a language to be extensional?

- Two expressions with the same extension (i.e., referring to the same thing) are interchangeable in any context.
- Example:
  - Both 'the morning star' and 'the evening star' refer to Venus.
  - In the sentence "the morning star is the second planet from the sun" replacing 'the morning star' with 'the evening star' doesn't change the truth of the statement.

Extensional vs Intensional Language

What about intensional language?

- Intensional language involves contexts where substitution of identical terms changes meaning.
- Example:
  - John believes that the morning star is the brightest star in the morning sky.
  - John does not believe that the evening star is the brightest star in the morning sky.
  - Although 'morning star' and 'evening star' refer to the same object, they are not interchangeable in this context.

Argument against reduction/physicalism: Brentano

- The argument hinges on premise (4): Can definitions alone transform extensional physics-language into intensional mind-language?
- Strictly speaking, no:
  - Mental expressions differ for x and y even when $x = y$, while physical descriptions cannot capture this difference in an extensional framework.

Argument against reduction/physicalism: Brentano

- However, if we include syntax (a physical representation of mental concepts), we can distinguish intensional contexts in physics:
  - Write B(J, V, 'morning star', mbright) to mean: John believes of Venus, described as 'the morning star' that it is the brightest star in the morning sky.
  - This allows us to say:
    - B(J, V, 'morning star', mbright) holds.
    - But B(J, V, 'evening star', mbright) does not.
- A risk of this approach: too many distinctions.
  - John should believe that a couch is comfy whether it's called 'couch' or 'sofa'.

Reconciling Reduction: Supervenience

- A possible reply to the limits of reductionism is to propose weaker forms of reduction between the mental and the physical.
- Instead of claiming that the mental is the physical, we can argue that the mental is determined by the physical while remaining distinct and described in its own intensional language.
- Supervenience is a philosophical concept that captures this weaker relationship.
- To say that the mental supervenes on the physical means:
  - There can be no difference in mental properties without a corresponding difference in physical properties.
  - Example: It's impossible for two agents to have identical brain, body, and environmental states, but differ in their beliefs or mental states.

Reconciling Reduction: Supervenience

- Supervenience describes a minimal form of dependence:
  - The mental depends on the physical, but the two can still be of fundamentally different kinds.
  - This avoids full reductionism while preserving the link between the mental and the physical.
- Although supervenience avoids strict reductionism, it doesn't explain why mental properties arise from physical ones.

Outline

Internalism vs. Externalism

- A key question about mental states: When do two agents have the same mental states?
- Two views:
  - Internalism: Mental states depend only on what is internal to the agent (e.g., brain states, cognitive processes).
  - Externalism: Mental states also depend on the agent's environment and interactions with the world.

Internalism vs. Externalism

- Intuitively, internalism feels correct-mental states seem private and tied to what's "in the head".
- However, Putnam challenges this intuition with his famous twin earth argument.

Twin Earth argument

- Imagine a twin earth, identical to Earth in every way except that what looks and behaves like water there isn't H_2O but XYZ.
- Now suppose:
  - I desire water (the substance that is H_2O) on Earth.
  - My twin on twin earth, who is identical to me internally, desires what they call "water" (the substance that is XYZ).
- Are we in the same mental state? Arguably not.
  - The content of my mental state is tied to H_2O.
  - The content of my twin's mental state is tied to XYZ.
- Hence, mental states depend on the external environment, not just internal brain states, supporting externalism.

Twin Earth argument

- If externalism is correct, mental states don't supervene on just physical brain states-they also depend on the environment.
- Mental states may have (at least) two components:
  - An internal component (narrow content): Internal representations of concepts (e.g., the concept of "water").
  - An external component (wide content): The semantic relationship between these internal representations and their referents in the world (e.g., H_2O or XYZ).

Relevance to AI

- Internalism and externalism raise key questions for AI:
  - Do AI systems represent "internal mental states" or are their states entirely dependent on external programming and data?
  - If externalism applies, an AI system's "knowledge" or "desires" would depend not just on its internal architecture, but also on the environment it interacts with.

Methodology: Thought Experiments

- In philosophy, thought experiments are one of the most useful tools for exploring abstract concepts and testing arguments.
- We've encountered several thought experiments so far:
  - Descartes' Demon: Can we trust our perceptions of reality?
  - Brain-in-a-vat: What defines consciousness or experience?
  - Ship of Theseus: What constitutes identity over time?
  - Twin Earth: Are mental states determined internally or externally?
- However, thought experiments require careful analysis.
  - What kind of possibility is the imagined scenario supposed to represent?

Methodology: Thought Experiments

- Logical Possibility:
  - A scenario is logically possible if it doesn't involve a contradiction.
  - Example: Descartes' Demon demonstrates that it is logically possible for all of our perceptions to be deceived, challenging whether observations justify belief in a real world.
- Physical Possibility:
  - A scenario is physically possible if it conforms to the laws of nature.
  - Example: Imagining two physically identical agents with different mental states is problematic because physical laws may not allow for such cases.
- Logical possibility is easier to establish but often less relevant to real-world questions. Physical possibility is harder to prove but more impactful for debates like physicalism.

Methodology: Thought Experiments

- Thought experiments are crucial for testing the boundaries of AI capabilities and ethical frameworks:
  - Brain-in-a-vat parallels questions about virtual reality and simulated experiences in AI.
  - Twin Earth applies to AI systems trained in different environments: Does their "understanding" depend on the data they've been exposed to?
  - Ship of Theseus raises questions about AI identity: If we replace all components of an AI, is it the same system?
- Distinguishing logical and physical possibilities in AI:
  - Logically possible: Could an AI simulate human consciousness entirely through algorithms?
  - Physically possible: Can such an AI be built with current or future technology?

Outline

Functionalism

- Functionalism: mental states are determined by their functional roles rather than by their physical or material composition.
- A mental state is identified by:
  - its causal connections to sensory inputs (e.g., perceiving light or sound),
  - its interactions with other mental states (e.g., forming beliefs or desires), and
  - its role in producing behavior (e.g., moving toward a stimulus).
- Thus, it doesn't matter how a mental state is realized (whether in a human brain, a silicon chip, or an alien organism)-what matters is what it does.

Functionalism

- Key implication: Functionalism supports the idea of multiple realizability:
  - The same mental state (e.g., pain) could exist in different physical systems, as long as the system plays the same functional role.
  - For example, humans, animals, and potentially AI could all experience "pain" if their systems function equivalently.

Functionalism

- Functionalism aligns closely with how AI systems are designed:
  - AI systems simulate mental states (e.g., reasoning and decision-making) by replicating the functional processes of the mind.
  - They do not require a biological brain; what matters is that they perform the same roles.
- Examples:
  - A chatbot "believes" it should provide relevant answers by processing input, accessing a database, and generating output- mirroring the functional role of belief.
  - Self-driving cars "decide" to stop at a red light by integrating sensor data and executing control algorithms, analogous to how humans use perception and reasoning to act.

Challenges

- Critics argue that functionalism may not fully capture the qualitative (or phenomenal) aspects of mental states-what it is like to experience them.
- Thought experiments (e.g., Searle's Chinese Room, the inverted spectrum) challenge whether a purely functional description can account for subjective experience.
- More on this tomorrow!

Exercises

1. Compare the extensional vs. intensional distinction of languages to programming languages that are functional vs. having side effects. What are commonalities, what are differences?
2. Can you think of more kinds of possibility other than 'logical' and 'physical'? Which notions of possibility are used in the indistinguishably experience argument scheme?
3. Look back at some of the arguments (Bretano, Twin Earth, ...): analyze them carefully, where would you object? Keep in mind the discussion on the methodology of using thought experiments.