

Philosophy of Consciousness and AI

Amine Rusi El Hassani

Philosophical perspectives on the exact sciences I

October 1, 2025

Two Questions

- What is consciousness?
- Can AI become conscious?

Today's Journey

- ① Core philosophical positions on consciousness (Dualism, Physicalism etc.).
- ② Central debates on AI consciousness

Can AI Become Conscious?

- This central question combines two parts:
 - ① **What does it mean to be conscious?** – To be *awake to the world*, to feel that there is *something it is like* to be that entity.

Can AI Become Conscious?

- This central question combines two parts:
 - ① **What does it mean to be conscious?** – To be *awake to the world*, to feel that there is *something it is like* to be that entity.
 - ② **Could an artificial intelligence possess this?** – Not just *simulate* intelligence or behavior, but genuinely have an inner perspective.

Can AI Become Conscious?

- This central question combines two parts:
 - ① **What does it mean to be conscious?** – To be *awake to the world*, to feel that there is *something it is like* to be that entity.
 - ② **Could an artificial intelligence possess this?** – Not just *simulate* intelligence or behavior, but genuinely have an inner perspective.
- Consciousness is not just problem-solving or output generation. It is about the **qualitative feel** of experience (seeing red, feeling pain, enjoying music).

Can AI Become Conscious?

- This central question combines two parts:
 - ① **What does it mean to be conscious?** – To be *awake to the world*, to feel that there is *something it is like* to be that entity.
 - ② **Could an artificial intelligence possess this?** – Not just *simulate* intelligence or behavior, but genuinely have an inner perspective.
- Consciousness is not just problem-solving or output generation. It is about the **qualitative feel** of experience (seeing red, feeling pain, enjoying music).
- Raises deep issues:
 - How do we define and detect consciousness?
 - Can purely physical processes ever give rise to subjective experience?
 - If a machine behaves *as if* conscious, is that enough?

Clarifying Terms

Consciousness

- access consciousness (**A**-consciousness): information globally available for report and control.
- *phenomenal* consciousness (**P**-consciousness): the felt, subjective *what-it-is-like* experience.

Clarifying Terms

Consciousness

- access consciousness (**A**-consciousness): information globally available for report and control.
- *phenomenal* consciousness (**P**-consciousness): the felt, subjective *what-it-is-like* experience.

Bridge: understanding **A** is scientifically valuable, but the hard problem (why it feels like anything) **is still unsolved**. Is a *thermostat* conscious? A *mouse*? Where would *you* draw the line?

Why Scientists Should Care

- AI and neuroscience deal with **access (A)**: how information is processed, shared, reported.

Why Scientists Should Care

- AI and neuroscience deal with **access (A)**: how information is processed, shared, reported.
- But the AI consciousness debate hinges on: does access = real consciousness, or just its appearance?

Why Scientists Should Care

- AI and neuroscience deal with **access (A)**: how information is processed, shared, reported.
- But the AI consciousness debate hinges on: does access = real consciousness, or just its appearance?
- Philosophy helps to clarify the **assumptions** behind interpreting data and models.

Why Scientists Should Care

- AI and neuroscience deal with **access (A)**: how information is processed, shared, reported.
- But the AI consciousness debate hinges on: does access = real consciousness, or just its appearance?
- Philosophy helps to clarify the **assumptions** behind interpreting data and models.
- Without philosophy, science risks mistaking **acting conscious** for **being conscious**.

Acting Conscious vs. Being Conscious

- **Acting conscious:**

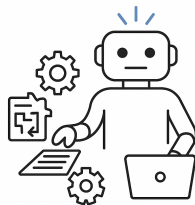
- Producing human-like behavior
- Passing language tests, solving problems
- Simulation of intelligence

- **Being conscious:**

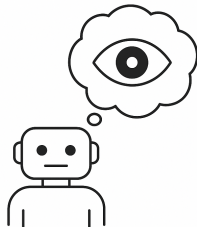
- Having subjective experience
- Inner life: qualia, *something it is like*
- First-person perspective

- The challenge: Does perfect simulation imply real consciousness?

Acting Conscious



Being Conscious



Acting vs. Being: function vs. experience

Key Arguments: The Hard Problem of Consciousness

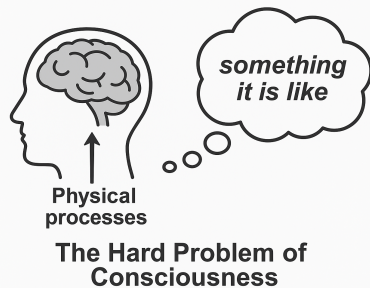
- **Easy problems:** explaining cognitive and behavioral functions
 - Perception, attention, memory, learning
 - Information processing and control of behavior

Key Arguments: The Hard Problem of Consciousness

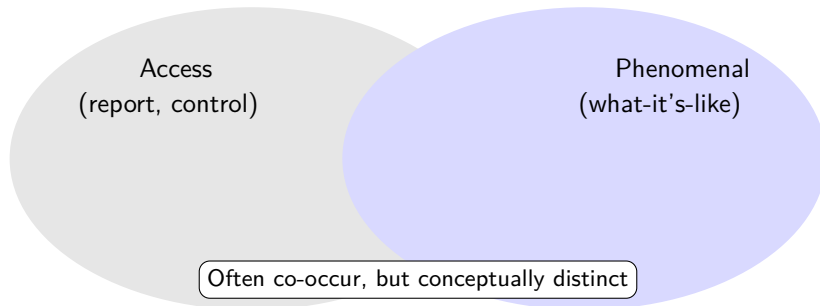
- **Easy problems:** explaining cognitive and behavioral functions
 - Perception, attention, memory, learning
 - Information processing and control of behavior
- **Hard problem** (Chalmers 1995/1996):
 - Why is there *something it is like* to undergo these processes?
 - How and why do physical processes give rise to **qualia** — felt experiences such as seeing red, feeling pain, or tasting coffee?

Key Arguments: The Hard Problem of Consciousness

- **Easy problems:** explaining cognitive and behavioral functions
 - Perception, attention, memory, learning
 - Information processing and control of behavior
- **Hard problem** (Chalmers 1995/1996):
 - Why is there *something it is like* to undergo these processes?
 - How and why do physical processes give rise to **qualia** — felt experiences such as seeing red, feeling pain, or tasting coffee?
- Sets the stage for debates about whether AI could ever be *truly conscious*.



Access vs Phenomenal



Neuroscience models mostly target *access*; the hard problem concerns *phenomenal* feel.

What is consciousness?

Everything

Panpsychism

Something

*Physicalism
Dualism*

Nothing

*Illusionism
Eliminativism*

Panpsychism — Plain Idea

Core idea

- Consciousness is a basic feature of the physical world.
- Even simple physical systems (atoms, particles) have tiny elements of experience.
- Our human consciousness comes from the combination and organization of these micro-experiences.

Panpsychism — Plain Idea

Core idea

- Consciousness is a basic feature of the physical world.
- Even simple physical systems (atoms, particles) have tiny elements of experience.
- Our human consciousness comes from the combination and organization of these micro-experiences.

Panpsychism — Plain Idea

Core idea

- Consciousness is a basic feature of the physical world.
- Even simple physical systems (atoms, particles) have tiny elements of experience.
- Our human consciousness comes from the combination and organization of these micro-experiences.

Why consider it?

- It avoids the “leap” from pure physics to rich experience — experience is there all along.

Panpsychism — Plain Idea

Core idea

- Consciousness is a basic feature of the physical world.
- Even simple physical systems (atoms, particles) have tiny elements of experience.
- Our human consciousness comes from the combination and organization of these micro-experiences.

Why consider it?

- It avoids the “leap” from pure physics to rich experience — experience is there all along.
- Explains why physical science gives us structure and behavior, but not how it *feels*.

Panpsychism — Plain Idea

Core idea

- Consciousness is a basic feature of the physical world.
- Even simple physical systems (atoms, particles) have tiny elements of experience.
- Our human consciousness comes from the combination and organization of these micro-experiences.

Why consider it?

- It avoids the “leap” from pure physics to rich experience — experience is there all along.
- Explains why physical science gives us structure and behavior, but not how it *feels*.
- Offers a middle way: not dualism (separate mind stuff), not reductionism (nothing but physics).

Panpsychism — Plain Idea

Core idea

- Consciousness is a basic feature of the physical world.
- Even simple physical systems (atoms, particles) have tiny elements of experience.
- Our human consciousness comes from the combination and organization of these micro-experiences.

Why consider it?

- It avoids the “leap” from pure physics to rich experience — experience is there all along.
- Explains why physical science gives us structure and behavior, but not how it *feels*.
- Offers a middle way: not dualism (separate mind stuff), not reductionism (nothing but physics).

Who has this view?

Panpsychism — Main Doubts

- **Combination problem:** How do many tiny “micro-experiences” add up to one unified mind?

Panpsychism — Main Doubts

- **Combination problem:** How do many tiny “micro-experiences” add up to one unified mind?
- **Counter-intuitive:** It seems strange to say that an electron “feels” anything.

Panpsychism — Main Doubts

- **Combination problem:** How do many tiny “micro-experiences” add up to one unified mind?
- **Counter-intuitive:** It seems strange to say that an electron “feels” anything.
- **No direct evidence:** The theory is metaphysical; hard to test scientifically.

Panpsychism — Main Doubts

- **Combination problem:** How do many tiny “micro-experiences” add up to one unified mind?
- **Counter-intuitive:** It seems strange to say that an electron “feels” anything.
- **No direct evidence:** The theory is metaphysical; hard to test scientifically.
- **Explanatory risk:** It may push the mystery down a level rather than solve it.

Panpsychism — Main Doubts

- **Combination problem:** How do many tiny “micro-experiences” add up to one unified mind?
- **Counter-intuitive:** It seems strange to say that an electron “feels” anything.
- **No direct evidence:** The theory is metaphysical; hard to test scientifically.
- **Explanatory risk:** It may push the mystery down a level rather than solve it.

Typical replies

- Physics itself only tells us about relations and structures.

Panpsychism — Main Doubts

- **Combination problem:** How do many tiny “micro-experiences” add up to one unified mind?
- **Counter-intuitive:** It seems strange to say that an electron “feels” anything.
- **No direct evidence:** The theory is metaphysical; hard to test scientifically.
- **Explanatory risk:** It may push the mystery down a level rather than solve it.

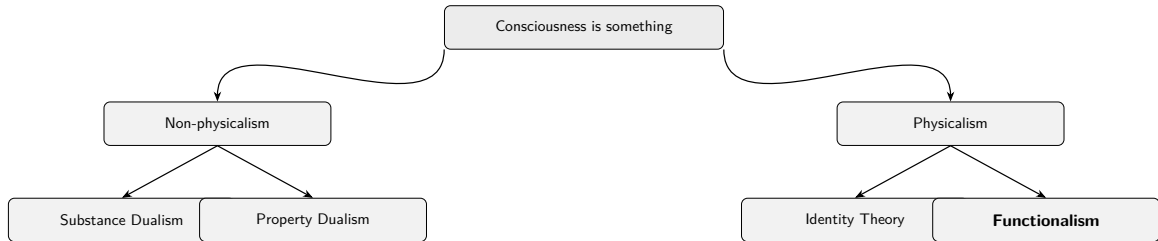
Typical replies

- Physics itself only tells us about relations and structures.
- Counter-intuitiveness is not decisive — many accepted theories (relativity, quantum mechanics) are counter-intuitive too.

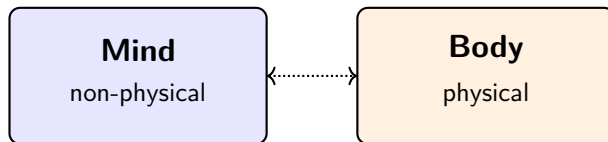
If Panpsychism is True — What About AI?

- Any system made of matter may already carry some form of consciousness.
- The question becomes: how much, and how organized?
- An AI made of silicon might have its own micro-conscious elements.
- "True consciousness" (one similar to human's) would depend on how these are combined and integrated.

If consciousness is *something*, what is it?



Substance Dualism — Diagram



Two distinct kinds of *substance*; they can interact but are not the same.

Substance Dualism — Plain Idea

Simple idea

- There are two kinds of things: **mind** and **matter**.
- The **mind** is the "I" that feels and thinks.
- The **brain** is physical machinery.
- They can affect each other, but they are not the same.

Substance Dualism — Plain Idea

Simple idea

- There are two kinds of things: **mind** and **matter**.
- The **mind** is the "I" that feels and thinks.
- The **brain** is physical machinery.
- They can affect each other, but they are not the same.

Why take it seriously?

- **One self:** My experiences belong to a single "I"; matter is made of parts (Descartes).

Substance Dualism — Plain Idea

Simple idea

- There are two kinds of things: **mind** and **matter**.
- The **mind** is the "I" that feels and thinks.
- The **brain** is physical machinery.
- They can affect each other, but they are not the same.

Why take it seriously?

- **One self:** My experiences belong to a single "I"; matter is made of parts (Descartes).
- **I without this body?:** It seems possible that "I" continue to exist even if this body does not.

Substance Dualism — Plain Idea

Simple idea

- There are two kinds of things: **mind** and **matter**.
- The **mind** is the "I" that feels and thinks.
- The **brain** is physical machinery.
- They can affect each other, but they are not the same.

Why take it seriously?

- **One self:** My experiences belong to a single "I"; matter is made of parts (Descartes).
- **I without this body?:** It seems possible that "I" continue to exist even if this body does not. **Who has this view?**

Substance Dualism — Main Doubts

- **Interaction:** How can a non-physical mind move neurons?

Substance Dualism — Main Doubts

- **Interaction:** How can a non-physical mind move neurons?
- **Brain causes:** Every brain event already seems explained by physical causes (neurons and chemistry). Where does mind fit?

Substance Dualism — Main Doubts

- **Interaction:** How can a non-physical mind move neurons?
- **Brain causes:** Every brain event already seems explained by physical causes (neurons and chemistry). Where does mind fit?
- **Simplicity:** Adding a new kind of thing makes the picture more complex.

Substance Dualism — Main Doubts

- **Interaction:** How can a non-physical mind move neurons?
- **Brain causes:** Every brain event already seems explained by physical causes (neurons and chemistry). Where does mind fit?
- **Simplicity:** Adding a new kind of thing makes the picture more complex.
- **Neuroscience:** Brain damage and stimulation change experience in systematic ways.

Substance Dualism — Main Doubts

- **Interaction:** How can a non-physical mind move neurons?
- **Brain causes:** Every brain event already seems explained by physical causes (neurons and chemistry). Where does mind fit?
- **Simplicity:** Adding a new kind of thing makes the picture more complex.
- **Neuroscience:** Brain damage and stimulation change experience in systematic ways.

Possible dualist replies

- There may be basic bridge laws linking mind and brain.

Substance Dualism — Main Doubts

- **Interaction:** How can a non-physical mind move neurons?
- **Brain causes:** Every brain event already seems explained by physical causes (neurons and chemistry). Where does mind fit?
- **Simplicity:** Adding a new kind of thing makes the picture more complex.
- **Neuroscience:** Brain damage and stimulation change experience in systematic ways.

Possible dualist replies

- There may be basic bridge laws linking mind and brain.
- The assumption that “only physical causes exist” is an assumption of the scientific method, not a necessity about the world.

Substance Dualism — Main Doubts

- **Interaction:** How can a non-physical mind move neurons?
- **Brain causes:** Every brain event already seems explained by physical causes (neurons and chemistry). Where does mind fit?
- **Simplicity:** Adding a new kind of thing makes the picture more complex.
- **Neuroscience:** Brain damage and stimulation change experience in systematic ways.

Possible dualist replies

- There may be basic bridge laws linking mind and brain.
- The assumption that “only physical causes exist” is an assumption of the scientific method, not a necessity about the world.
- Simplicity matters, but not if it ignores how experience appears from the inside.

If Substance Dualism is True — What About AI?

- A system can act conscious without any inner life (behavior is not enough).

If Substance Dualism is True — What About AI?

- A system can act conscious without any inner life (behavior is not enough).
- More computation or larger models may not produce consciousness by itself.

If Substance Dualism is True — What About AI?

- A system can act conscious without any inner life (behavior is not enough).
- More computation or larger models may not produce consciousness by itself.
- A machine might be conscious only if a mind is somehow linked to it.

If Substance Dualism is True — What About AI?

- A system can act conscious without any inner life (behavior is not enough).
- More computation or larger models may not produce consciousness by itself.
- A machine might be conscious only if a mind is somehow linked to it.
- Testing would need more than output and behavior; we lack clear tests for the inner side.

Property Dualism — Plain Idea

Core idea

- There is only one kind of substance: the physical world.
- But this substance has two kinds of **properties**:
 - Physical properties (mass, charge, neural firing).
 - Mental properties (feeling pain, seeing red).
- Mental properties are **irreducible**: they cannot be explained away in purely physical terms.

Difference from Substance Dualism

- Substance Dualism: two kinds of *substances* (mind and matter).
- Property Dualism: one kind of *substance* (matter) with two kinds of **properties**.

Mary's Room Argument (Frank Jackson, 1982)

The Thought Experiment:

- Mary is a scientist who knows **all the physical facts** about color vision.



Mary's Room Argument (Frank Jackson, 1982)

The Thought Experiment:

- Mary is a scientist who knows **all the physical facts** about color vision.
- But she has lived in a black-and-white room her whole life.
- When she leaves and sees red for the first time, she **learns something new: what it is like** to see red.



Mary's Room Argument (Frank Jackson, 1982)



The Thought Experiment:

- Mary is a scientist who knows **all the physical facts** about color vision.
- But she has lived in a black-and-white room her whole life.
- When she leaves and sees red for the first time, she **learns something new**: **what it is like** to see red.

Philosophical point:

- Physical knowledge is not the whole story.
- There are facts about conscious experience (**qualia**) that go beyond physics.
- Supports **property dualism**: consciousness cannot be reduced to the physical.

Property Dualism — Motivations and Doubts

Why consider it?

- Keeps the scientific picture of a single physical world.

Property Dualism — Motivations and Doubts

Why consider it?

- Keeps the scientific picture of a single physical world.
- Still respects the “extra” character of consciousness (qualia, subjective feel).

Property Dualism — Motivations and Doubts

Why consider it?

- Keeps the scientific picture of a single physical world.
- Still respects the “extra” character of consciousness (qualia, subjective feel).
- Fits arguments like Mary’s Room: all physical facts are not all the facts.

Property Dualism — Motivations and Doubts

Why consider it?

- Keeps the scientific picture of a single physical world.
- Still respects the “extra” character of consciousness (qualia, subjective feel).
- Fits arguments like Mary’s Room: all physical facts are not all the facts.

Main doubts

- **Causal problem:** if mental properties are extra, how do they affect the brain?

Property Dualism — Motivations and Doubts

Why consider it?

- Keeps the scientific picture of a single physical world.
- Still respects the “extra” character of consciousness (qualia, subjective feel).
- Fits arguments like Mary’s Room: all physical facts are not all the facts.

Main doubts

- **Causal problem:** if mental properties are extra, how do they affect the brain?
- **Epiphenomenal worry:** mental properties might exist but do nothing.

Property Dualism — Motivations and Doubts

Why consider it?

- Keeps the scientific picture of a single physical world.
- Still respects the “extra” character of consciousness (qualia, subjective feel).
- Fits arguments like Mary’s Room: all physical facts are not all the facts.

Main doubts

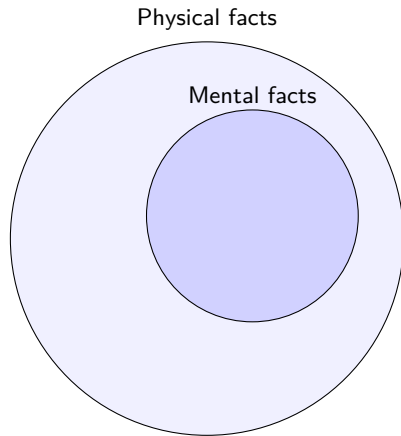
- **Causal problem:** if mental properties are extra, how do they affect the brain?
- **Epiphenomenal worry:** mental properties might exist but do nothing.
- **Explanatory risk:** says experience is “extra,” but gives no mechanism.

Property Dualism — Implications for AI

- Machines could have all the right **physical** and **functional** properties...
- ...but still lack the **mental properties** that give rise to felt experience.
- AI behavior could match humans perfectly without any inner life.
- Whether AI can be conscious depends on whether the right physical basis also carries the “extra” mental properties.

Difference from Substance Dualism

- Substance Dualism: AI would need a *mind-stuff*, unlikely in silicon.
- Property Dualism: AI could, in principle, have consciousness if physical systems like silicon also support mental properties.



Supervenience: no mental difference without a physical difference.

Physicalism — Core Idea

- Everything that exists is physical, or wholly depends on the physical.

Physicalism — Core Idea

- Everything that exists is physical, or wholly depends on the physical.
- If two worlds are the same in **all physical respects**, they are the same in **every respect**.

Physicalism — Core Idea

- Everything that exists is physical, or wholly depends on the physical.
- If two worlds are the same in **all physical respects**, they are the same in **every respect**.
- Mental states (like pain, seeing red) are not extra non-physical stuff; they are part of the physical story or depend on it.

Consciousness on a Physicalist View

- Experiences are either **identical** to, or **realized** by, neural processes.

Consciousness on a Physicalist View

- Experiences are either **identical** to, or **realized** by, neural processes.
- The **feel** of experience depends on **large-scale physical organization**.

Consciousness on a Physicalist View

- Experiences are either **identical** to, or **realized** by, neural processes.
- The **feel** of experience depends on **large-scale physical organization**.
- **No** extra non-physical properties are needed (although explanations may be incomplete today).

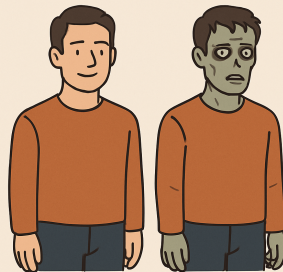
Famous Objections (Simple Version)

- **Explanatory gap:** knowing **mechanisms** seems **different** from knowing what **it is like**.
- **Zombie Twin argument:** a physical duplicate without experience is imaginable.

Zombie Twin

- **Idea:** a being physically/behaviorally just like us but with *no experience* (no “what it’s like”).
- **Point:** if such a duplicate is *conceivable*, then physical facts might not fix experiential facts.
- **Use against physicalism:** challenges “physical facts entail all facts.”
- **Common replies:**
 - Conceivable $\neq$ possible (concepts can mislead).
 - "Experience" is just a higher-level physical/functional pattern.
 - We lack the right bridging concepts/theory yet. But let's just keep looking for such theories!

Zombie twin:



Imagine a perfect physical copy that behaves like me but has no inner life

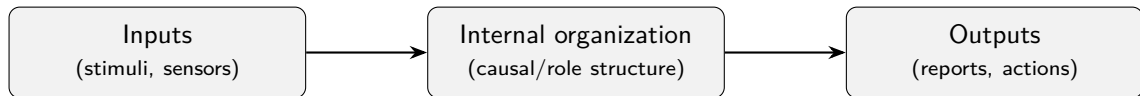
Varieties of Physicalism

- **Reductive identity:** a mental state is identical to a brain state-type.

Varieties of Physicalism

- **Reductive identity:** a **mental state** is **identical** to a **brain state-type**.
- **Non-reductive (realization):** **mental** properties are higher-level patterns **realized** by **physical** bases.
- **Example: Functionalism** mental states are defined by their **roles** (inputs, internal transitions, outputs) and are realized physically.

Functionalism (why it matters for AI)



Thesis: mental states = *functional roles* in the right organization.

Functionalism — The Core Idea

- A mental state = whatever **plays the right role**: links *inputs* (stimuli), *internal processing*, and *outputs* (behavior, reports).

Functionalism — The Core Idea

- A mental state = whatever **plays the right role**: links *inputs* (stimuli), *internal processing*, and *outputs* (behavior, reports).
- **Example: Pain** = the system's *damage-alarm* role (caused by injury, drives avoidance, teaches learning, prompts “ouch”).

Functionalism — The Core Idea

- A mental state = whatever **plays the right role**: links *inputs* (stimuli), *internal processing*, and *outputs* (behavior, reports).
- **Example: Pain** = the system's *damage-alarm* role (caused by injury, drives avoidance, teaches learning, prompts “ouch”).
- **Multiple realizability**: the same role could be played by human neurons, an octopus's nervous system, or (in principle) well-designed silicon.

Functionalism — Main Doubts

- **Feeling vs function:** Could something play the role of pain but with no pain *felt*?
- **Same behavior, no understanding?** (Chinese Room Argument): following rules may mimic meaning without any grasp.

The Chinese Room Argument

- **Setup:** A person who does not know Chinese is locked in a room.
- They receive Chinese symbols as input and consult a rulebook (in English).
- By following the rules, they produce appropriate Chinese responses as output.
- **Key Point:** From outside, it looks like understanding. Inside, there is only symbol manipulation.
- **Implication:** Computers may simulate understanding, but do not genuinely *understand*.



Realization: One Role, Many Implementations

- Think **software/app**: the *function* (what it does) can run on different hardware.

Realization: One Role, Many Implementations

- Think **software/app**: the *function* (what it does) can run on different hardware.
- In minds, the *functional pattern* (causal role) can be built from different physical parts.

Realization: One Role, Many Implementations

- Think **software/app**: the *function* (what it does) can run on different hardware.
- In minds, the *functional pattern* (causal role) can be built from different physical parts.
- So, mental kinds “ride on” physical stuff without being tied to a single neural micro-type.

Realization: One Role, Many Implementations

- Think **software/app**: the *function* (what it does) can run on different hardware.
- In minds, the *functional pattern* (causal role) can be built from different physical parts.
- So, mental kinds “ride on” physical stuff without being tied to a single neural micro-type.
- **A conscious AI** ? if the role/organization is sufficient, a suitably built machine *could* have mental states.

The Hardest Easy Question

