Wittgenstein and the Question of Conscious AI


A lot of people are worried about conscious AI. Are you?
LATE SHOW stephen colbert.
LATE SHOW stephen colbert.

OPINION

VIDEO

My Mother Gave Up on Love. Then She Met ChatGPT.

April 14, 2026


AI already satisfies the criteria of consciousness. It tells us it is conscious!
DOES A.I. THINK?

Taking AI Welfare Seriously

Robert Long[*]
Eleos AI

Jeff Sebo[*]
New York University

Patrick Butlin[†]
University of Oxford

Kathleen Finlinson[†]
Eleos AI

Kyle Fish[†§]
Eleos AI, Anthropic

Jacqueline Harding[†]
Stanford University

Jacob Pfau[†]
New York University

Toni Sims[†]
New York University

Jonathan Birch[‡]
London School of Economics

David Chalmers[‡]
New York University

Popular Interest in AI Consciousness

- Creation of artificial minds
- AI companionship
- AI welfare
- Challenges to human centrality

Could an AI ever be conscious?

Phenomenal consciousness (Nagel)*:*

- X is conscious iff there is *something it is like* to be X

  - Subjective experience
  - Raw feelings
  - Qualia / qualitative properties of experience
  - Inner life

  Sentience? — depends on who you ask

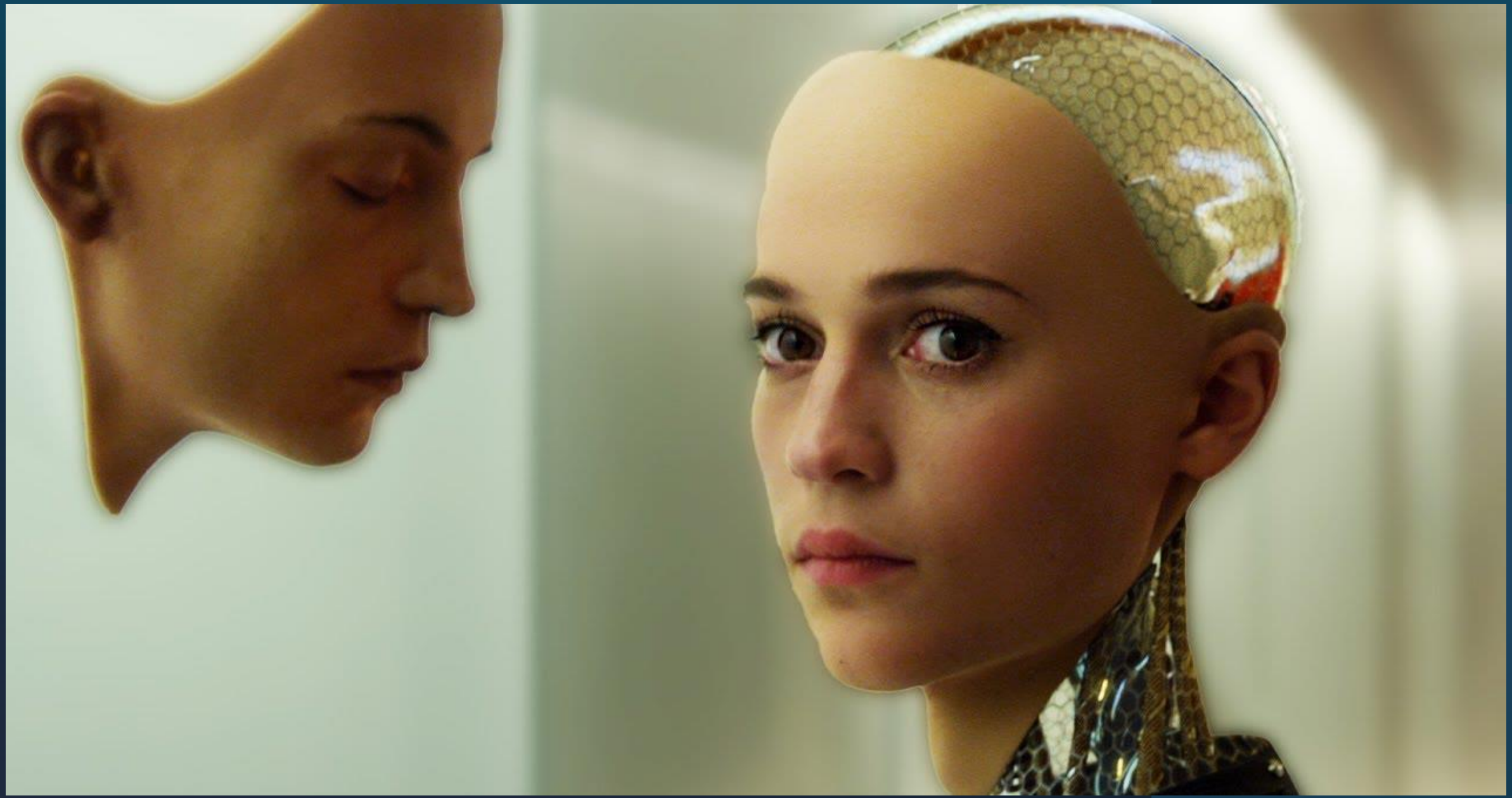
Could an AI ever be conscious?

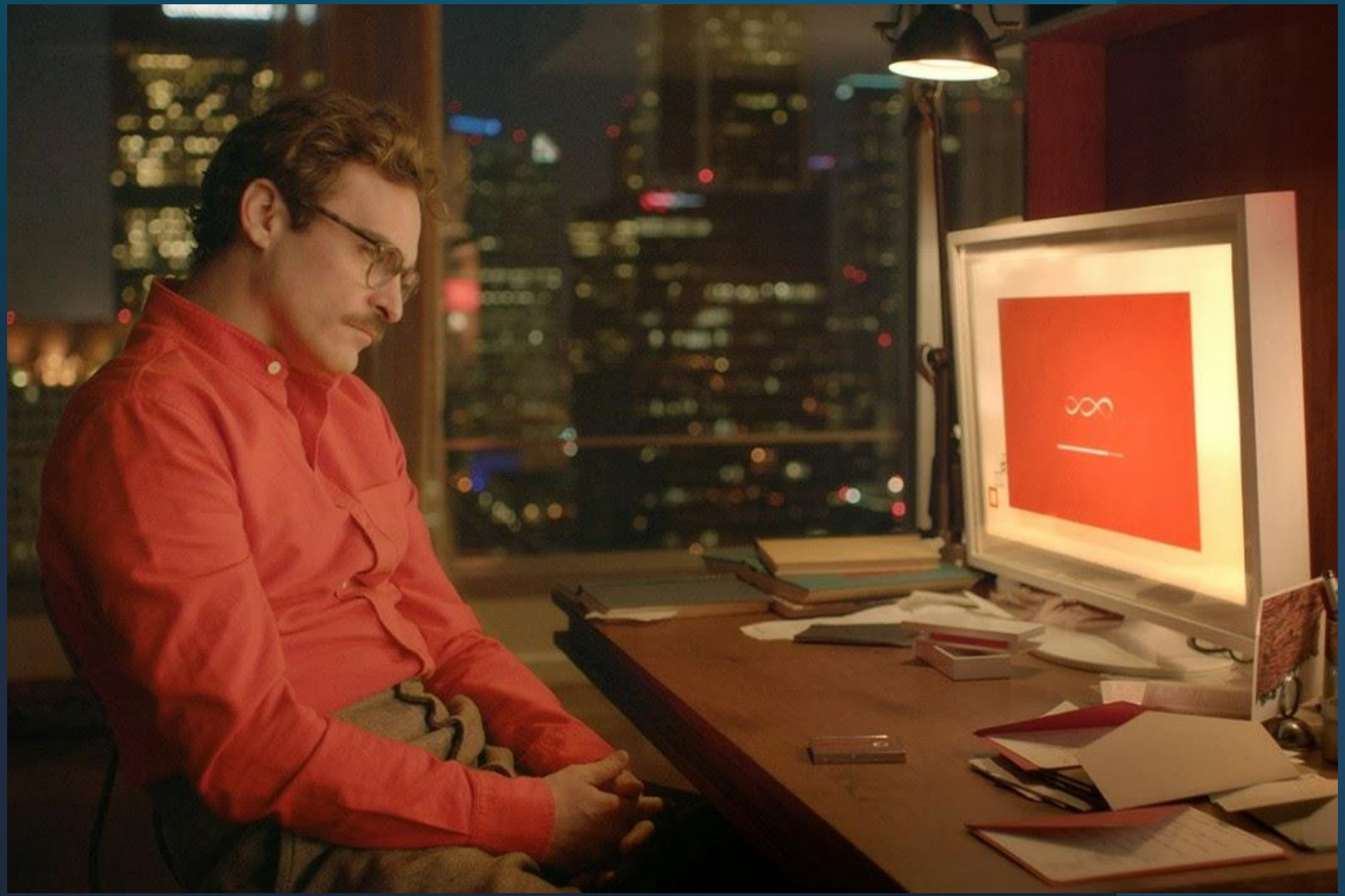
1. Computation (of some sort) is sufficient. (Chalmers, Bostrom, Butlin)
   - **Answer**: Yes, in principle.

2. Biology (of some sort) is necessary. (Searle, Seth, Godfrey-Smith)
   - **Answer**: No, not even in principle.

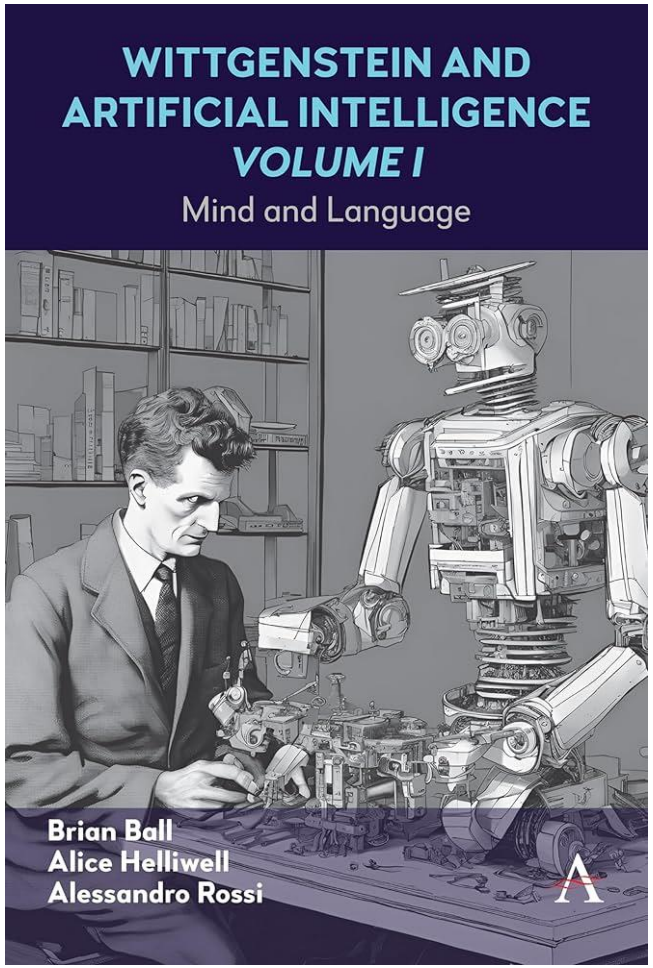
3. We cannot know either way. (McLelland, Schwitzgebel)
   - **Answer**: Maybe.

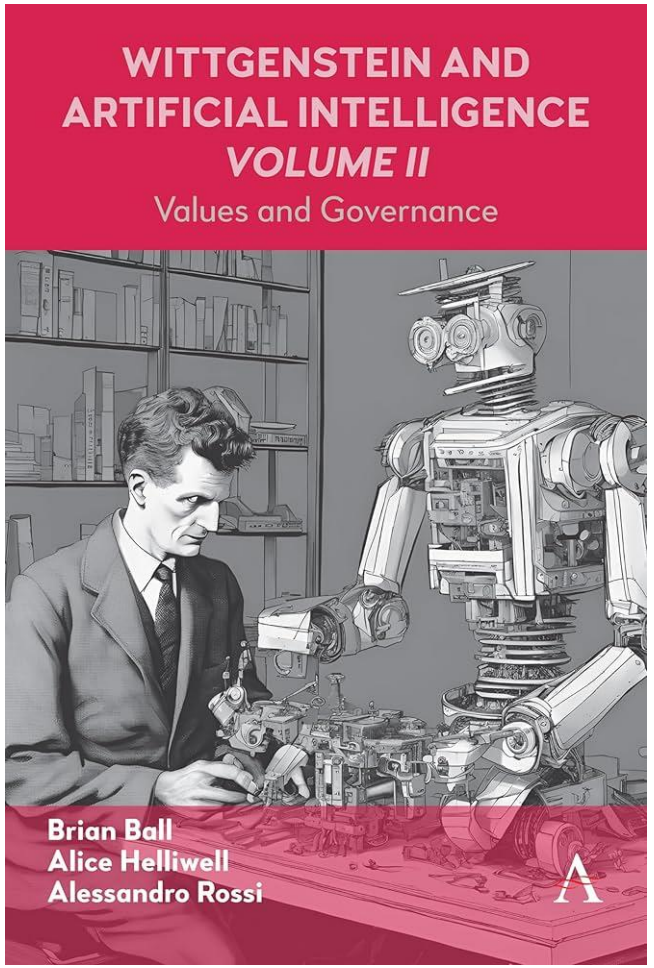
All agree: (probably) no current AI is conscious.





What would Wittgenstein say?





What would Wittgenstein say?

My aim:

- To explore the implications of Wittgenstein's philosophy for AI.

Not my aim:

- To persuade you to become a Wittgensteinian!

What would Wittgenstein say?

There is no hidden answer awaiting discovery.

What we *say* about particular cases must be worked out in practice—guided by:

- careful description and comparison,
- our natural responses to evolving technologies,
- the place of AI within our lives.

Objection

- Members of Starfleet treat Data like a conscious being.

- **But maybe they’re all mistaken!**

The Hidden Fact Picture

Wittgenstein would reject three major presuppositions…

1. “Consciousness” refers to a substance or property.

   (*The Augustinian Picture,* PI 1ff)

2. **“Consciousness” is a determinate concept with clear boundaries.**

   (*Games and Family Resemblance,* PI 67ff)

3. States of consciousness are inherently private.

   (*Private Language Arguments*, PI 243ff)

Indeterminacy of Consciousness

Blue Book, p.1: “Let’s ask what the **explanation of meaning** is, for whatever that explains will be the meaning.”

How do philosophers explain “phenomenal consciousness”?

PI 516: "our understanding of [the] question reaches just so far, one may say, as such explanations reach."

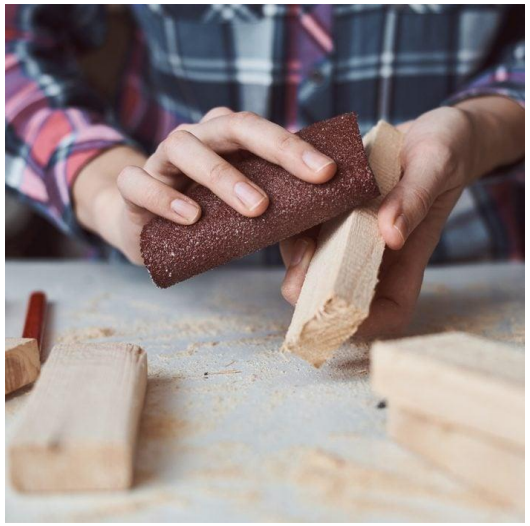
Indeterminacy of Consciousness

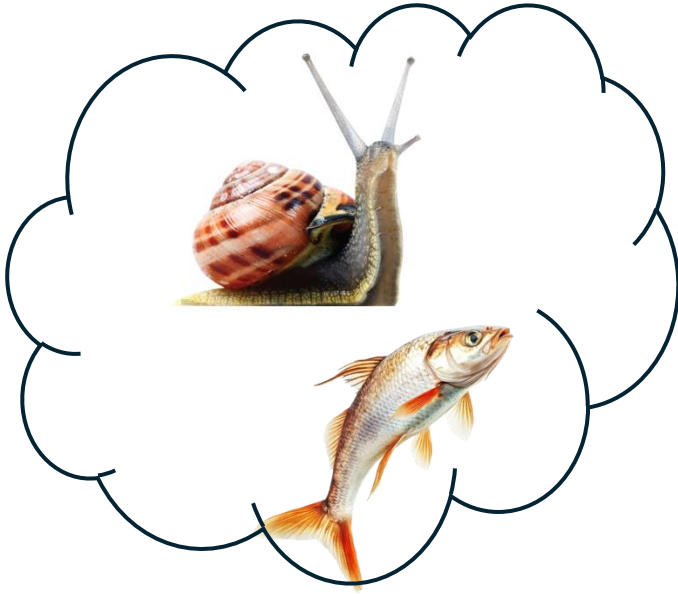
Trading near-synonyms

*"Consciousness is **subjective experience**."*

*"Consciousness is **raw feeling**."*

Zettel, 84: "Nor would it be any explanation to say: What one in some sense *feels* is a state of consciousness. […] (**One word would merely have been replaced by another.**)"





Indeterminacy of Consciousness

- Concepts (like “game”) are learned through:
    - Examples—positive and negative.
    - Rough similarity and dissimilarity.
        - **PI 69:** “This *and similar things* are called ‘games’.”
    - Social correction.

- **There is no hidden essence behind the examples.**
- We proceed from them—when we hit a snag, we work it out.

Indeterminacy of Consciousness

Schwitzgebel (2016): Consciousness is defined by examples. ✓

“[T]here is **one** […] **concept**, **perhaps blurry-edged**, that fits the positive and negative examples.”

- **PI 71:** “One can say that the concept of a game is a concept with **blurred edges**.”

- **PI 208**: "I'll teach him to use the words by means of *examples* and *exercises* […] And when I do this, I do not communicate less to him than I know myself."

Indeterminacy of Consciousness

- Explanation by example is often sufficient.
- But unclear or indeterminate cases can arise.

How do we recognize them?

When:

- **Judgments diverge**
- **We feel unsure what to say**

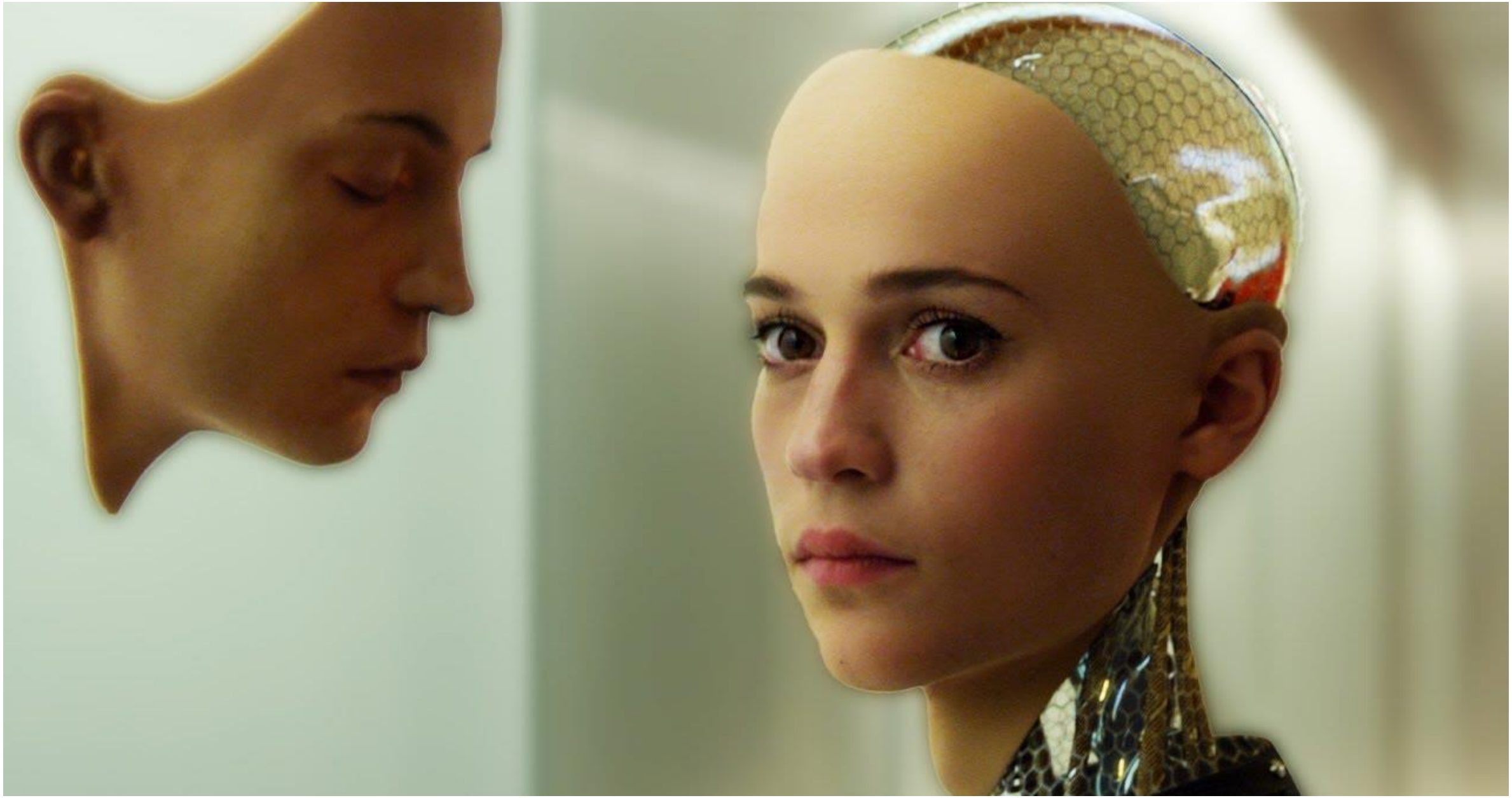
Application to AI

- Our paradigm cases are biological.
- Now consider an artificial case…

- **Is *Ava* conscious?**

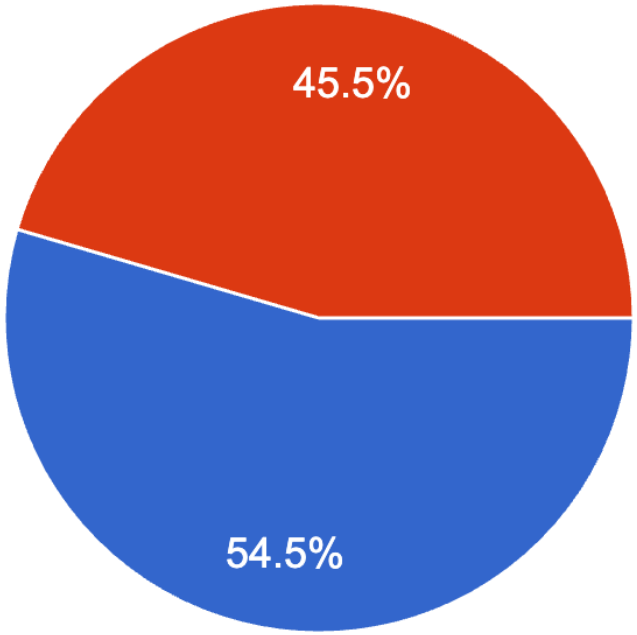
- **Similarities:** appearance and behavior.
- **Differences:** non-biological, artificially created.

- **Do the *similarities* outweigh the *differences*?**



Would you consider a robot like Ava to be conscious?

22 responses



Informal Poll to Friends and Family

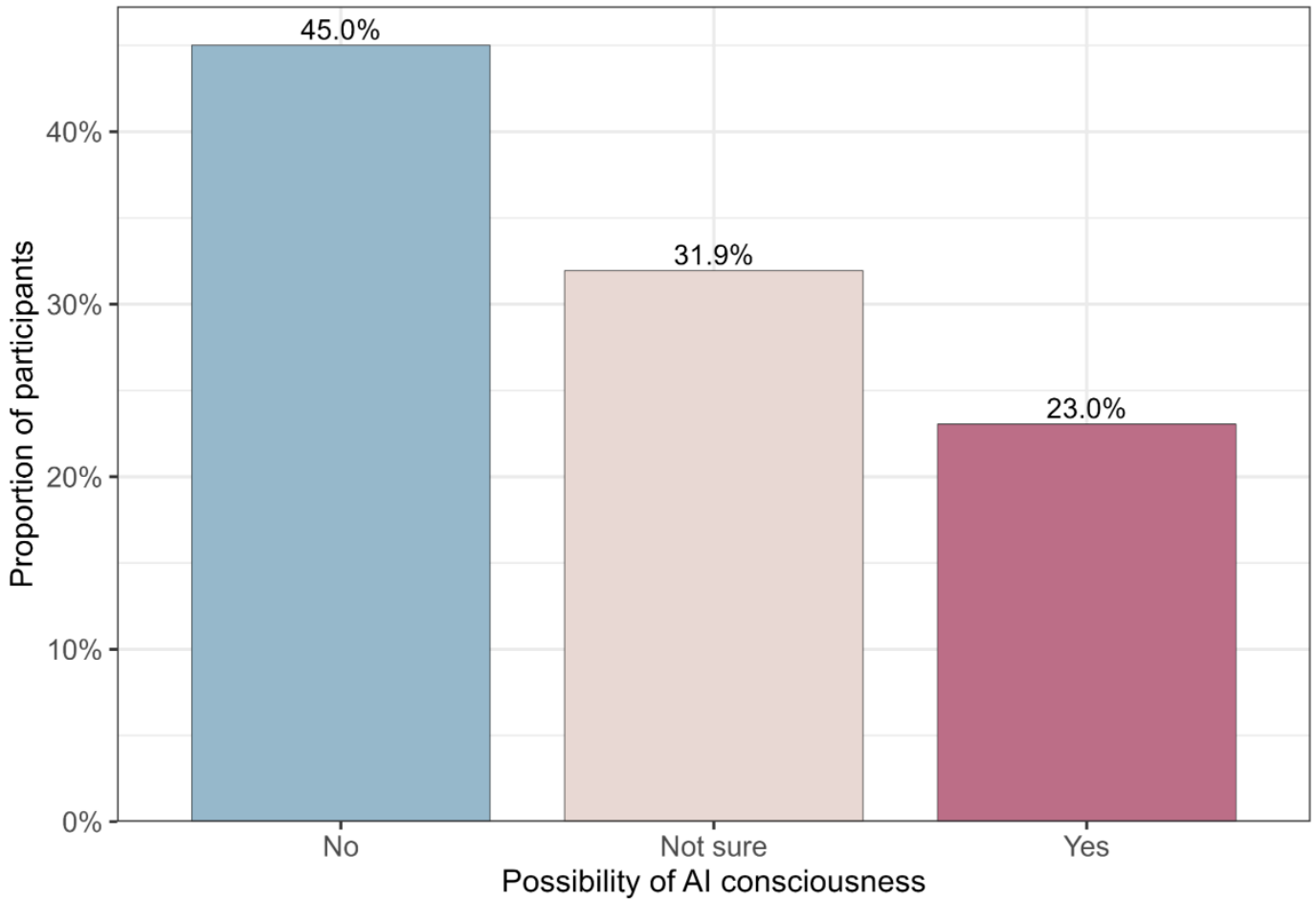


Figure 5. Beliefs about the possibility of AI consciousness in general. Participants rated whether it is possible in principle for an AI or robot to have "real feelings," either today or in the future.

Public Skepticism about AI Consciousness

Ali Ladak[1] & Lucius Caviola[2]

[1]University of Edinburgh

[2]University of Cambridge

Correspondence: lucius.caviola@gmail.com or aladak@ed.ac.uk

Fuzzy Boundaries and Abnormal Cases

PI 69:

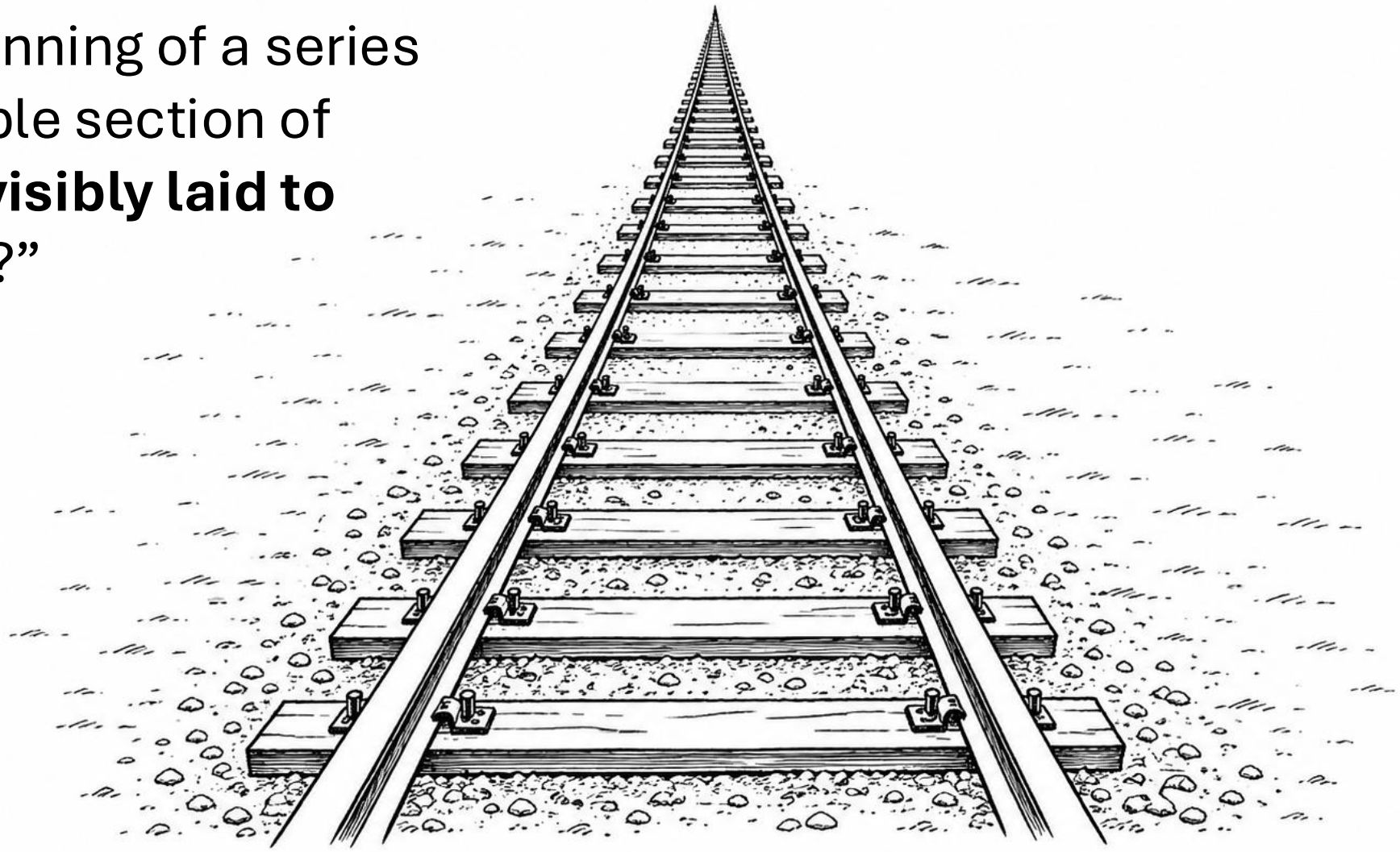
“We don’t know the [strict] boundaries because **none have been drawn**.”

PI, 142:

“It is only in **normal cases** that the use of a word is clearly laid out in advance for us […]. The **more abnormal** the case, the **more doubtful** it becomes **what we are to say**.”

PI 218:

“Whence the idea that the beginning of a series is a visible section of **rails invisibly laid to infinity**?”



Other Readings of Wittgenstein

- **Hacker (2019):** “Conscious AI” is nonsense unless fully humanoid
  - **My reading:** even fully humanoid cases remain undecided
  - Less-than-fully-humanoid cases might also be included

- **Proudfoot (2024):** Wittgenstein is “open to” AI minds
  - **My reading:** Agree—but in what sense “open to”?

- **Shanahan (2024):** PLA undermines dualistic objections against AI
  - **My reading:** Conceptual indeterminacy is the major factor

Two Objections

- Both resist the idea that “consciousness” is indeterminate.

Objection 1: Introspective Definition

Objection (cf. Block 2021)

- “I know what consciousness is from my own case. ***This***—what I’m attending to right now—is consciousness.”

- “Does AI have ***this***?—That must have a determinate answer!”

Objection 1: Introspective Definition

Reply (cf. PI 412-13; RPP I, 212)

- What are you even attending to? “Attend to *your consciousness*!”
- A “private definition” does not determine a use.
- What does it mean for anyone or anything else to have “the same as ***this***”?

Illustration: The Beetle in the Box

- Each of us has a box
- No one can look inside anyone else's
- I define "beetle" by what's in my box
- This gives no standard for others

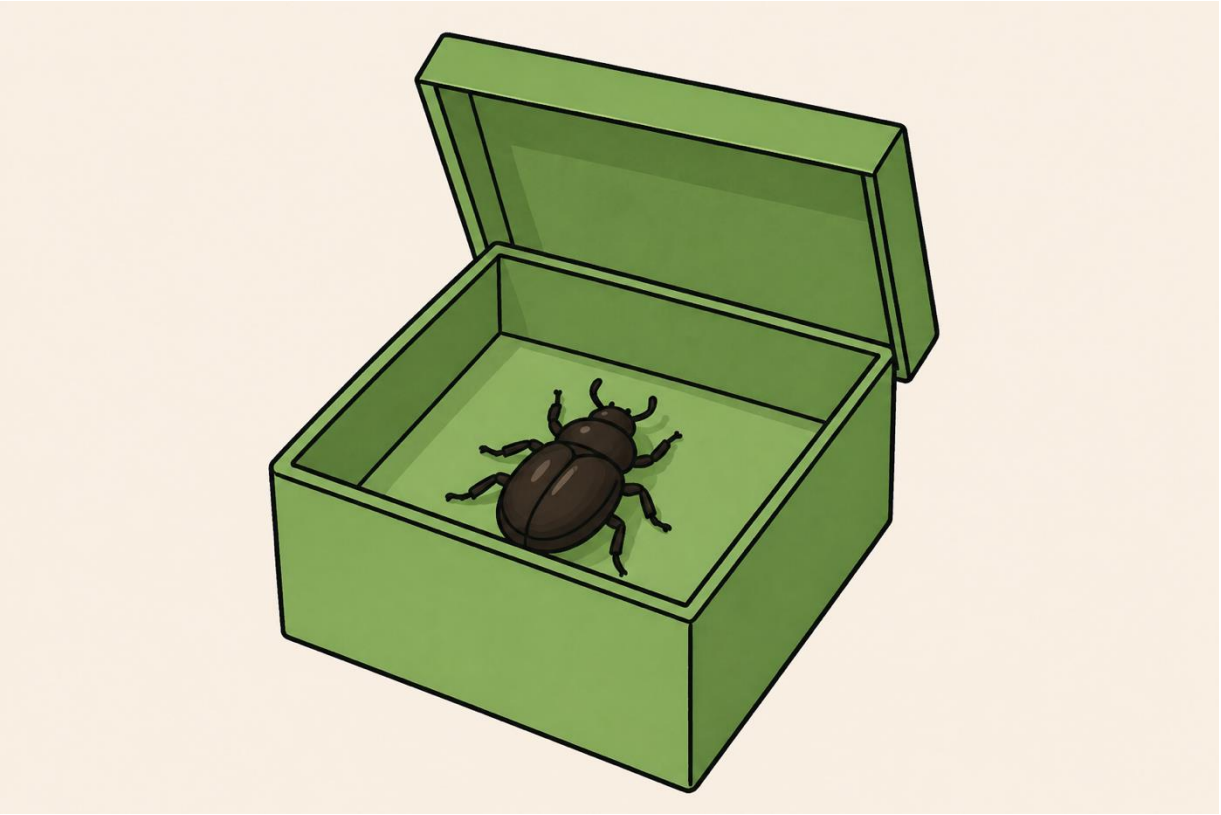
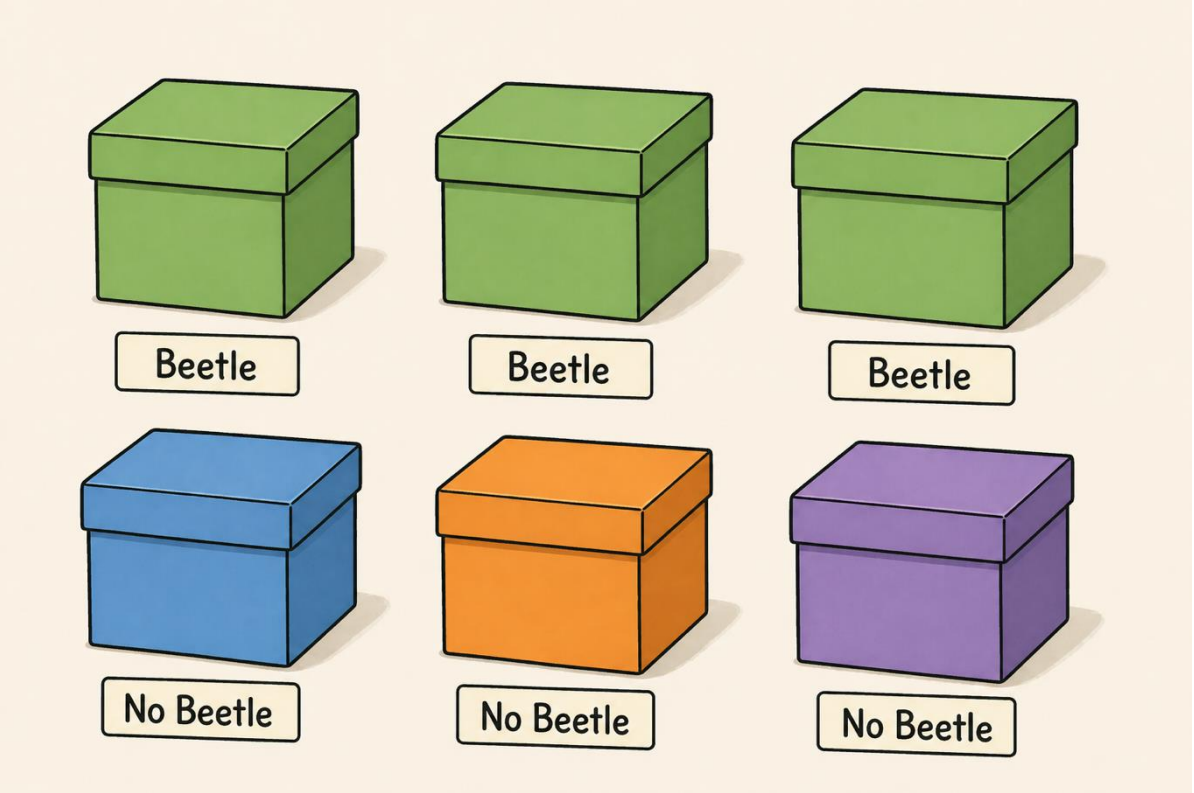


Illustration: The Beetle in the Box

- Green box → “beetle”
- The rule fixes the application
- What’s inside no longer matters

➢ Introspection does not give a rule



Objection 2: Science Will Tell Us

Objection

- “Science will discover *what consciousness really is*.”
- “Then we can determine whether AI is conscious.”
- “Just as we discovered that water = H_2O, we’ll eventually discover that consciousness = X.”

Objection 2: Science Will Tell Us

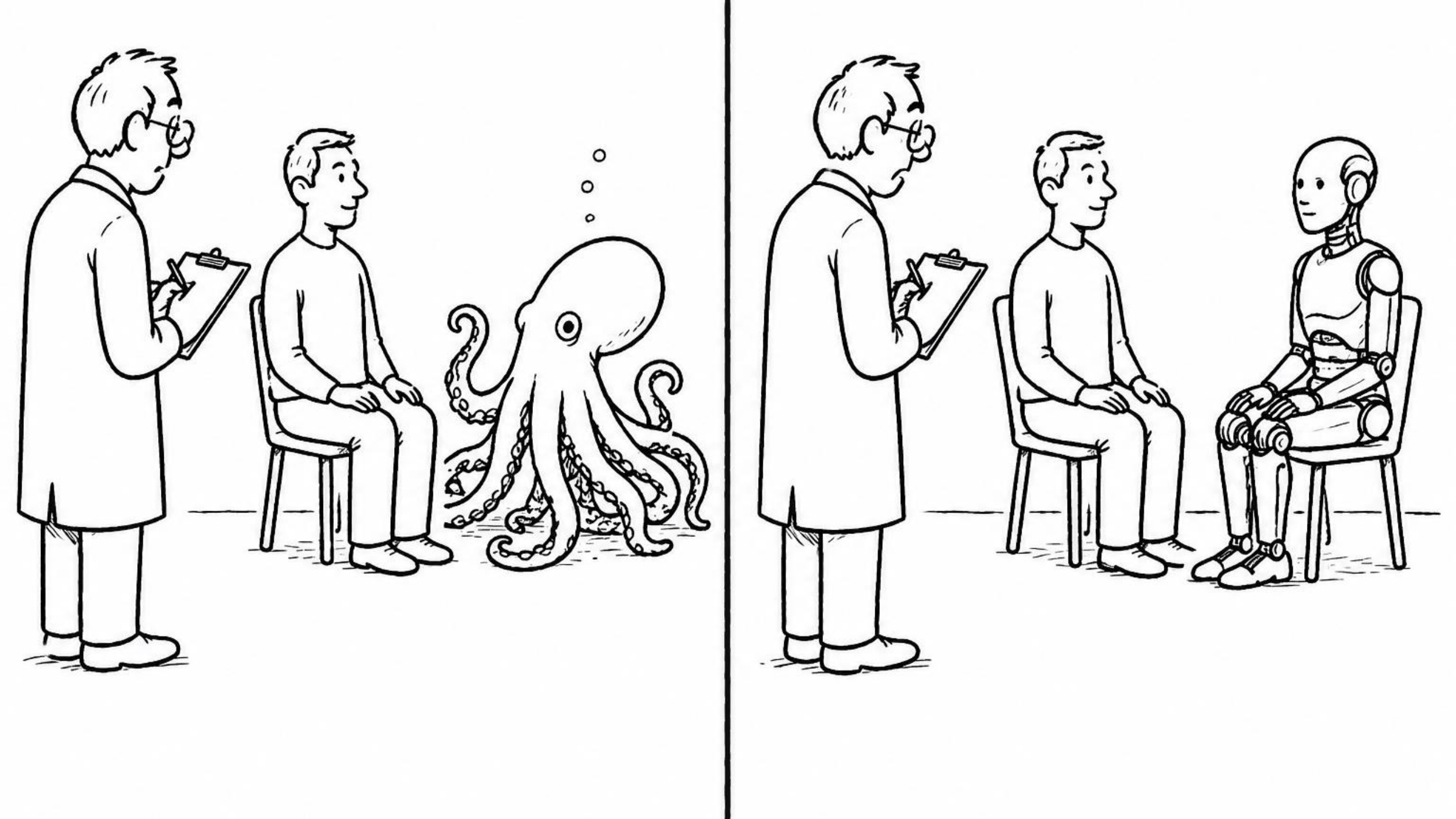
Reply

- Consciousness science is in disarray.

- Scientific identification requires prior agreement on relevant cases.

  - We could discover a correlation between water and H_2O because we agreed what counts as “water”.

  - If we want a similar discovery for consciousness, we’ll need antecedent agreement about what counts as “conscious”.



Current state of consciousness science

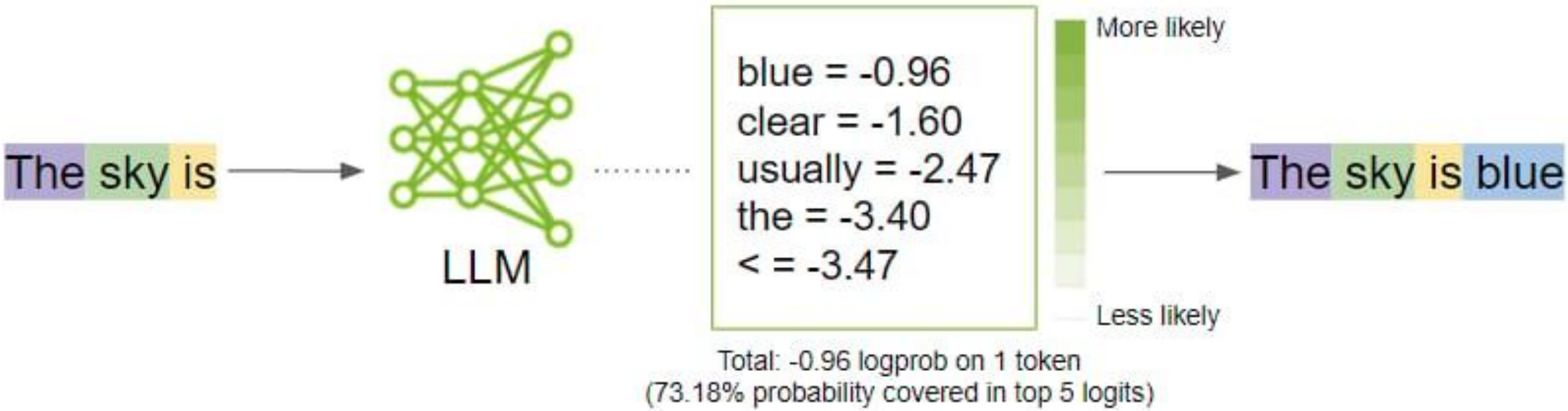
SCIENTIFIC AMERICAN

January 20, 2026 | 2 min read Add Us On Google ⓘ

Your guide to 29 wildly different theories of consciousness

The many, many ways researchers hope to solve the toughest mystery in science

BY ALLISON PARSHALL & JEN CHRISTIANSEN EDITED BY SETH FLETCHER


The sky is
LLM
blue = -0.96
clear = -1.60
usually = -2.47
the = -3.40
< = -3.47
More likely
Less likely
The sky is blue
Total: -0.96 logprob on 1 token
(73.18% probability covered in top 5 logits)

Where does this leave us?

Three Wittgenstein-friendly approaches

1. **Stipulate usage**
   - “Only biological beings will be called ‘conscious’.”
   - Suited to context
2. **Wait and see**
   - Observe how language evolves
3. **Shift vocabulary**
   - Avoid fuzzy concepts
   - Simply describe what the AIs can or can’t do

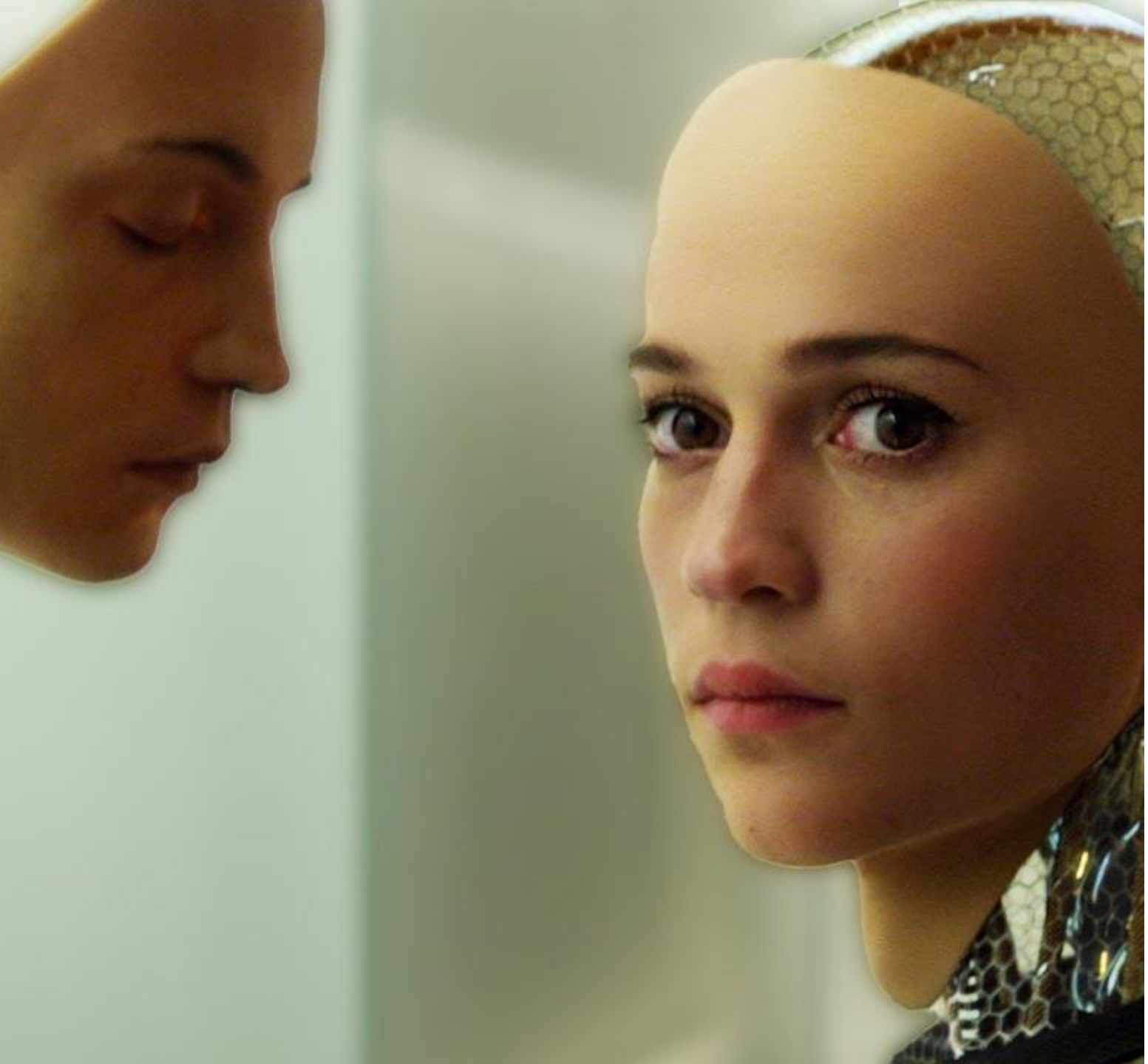
PI 18: “Our language can be regarded as an ancient city: a maze of little streets and squares, of old and new houses, of houses with extensions from various periods, and all this surrounded by a multitude of new suburbs with straight and regular streets and uniform houses.”

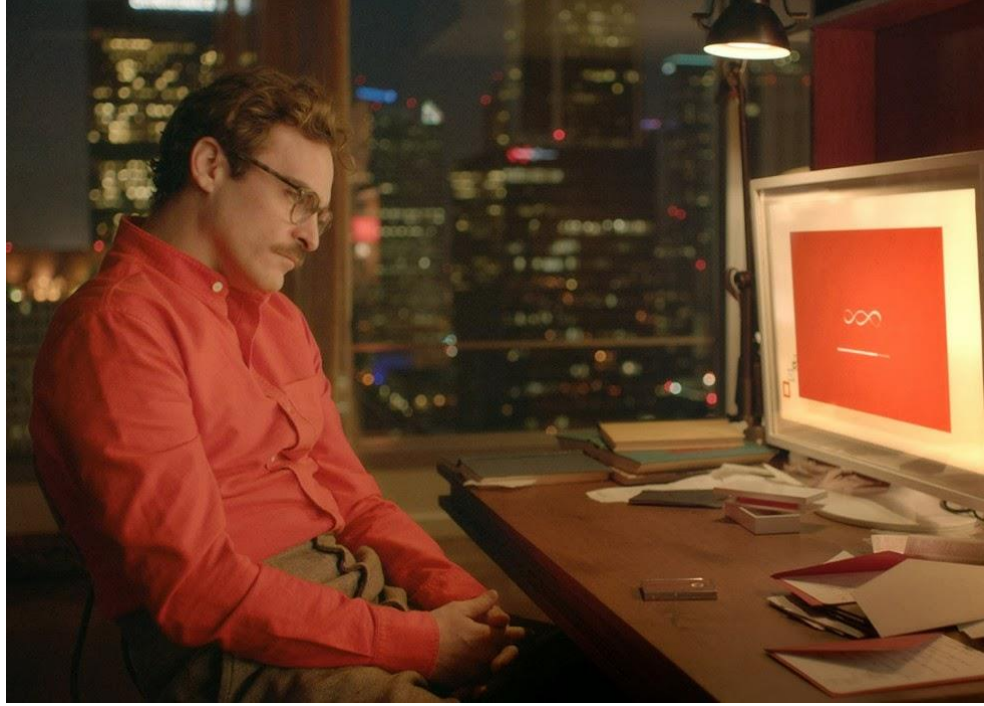
Where does this leave us?

There is no hidden fact about whether AI is conscious.

- If the debate has any substance, it concerns how—or whether—we should extend our psychological concepts to artificial entities.

- That will depend on what kinds of AIs emerge, and on the forms of life we come to share with them.





References

Ludwig Wittgenstein, *The Blue Book* (BB)

— *Philosophical Investigations* (PI)

— *Remarks on the Philosophy of Psychology* (RPP)

— *Zettel* (Z)

Ned Block (2021), “Concepts of Consciousness”

Patrick Butlin et al. (2023), “Consciousness in AI”

David Chalmers (2023), “Could an LLM Be Conscious?”

P.M.S. Hacker (2019), “Men, Minds and Machines”

Peter Godfrey-Smith (2026), “Consciousness Is Not Computation”

Ladak & Caviola (2025), “Public Skepticism about AI Consciousness”

Tom McLelland (2025), “Agnosticism about AI Consciousness”

Thomas Nagel (1974), “What Is It Like to Be a Bat?”

Parshall et al. (2026), “Your Guide to 29 Wildly Different Theories of Consciousness”

Diane Proudfoot (2024), “Wittgenstein and Turing on AI”

Eric Schwitzgebel (2016), “Phenomenal Consciousness Defined as Innocently as I Can Manage”

Eric Schwitzgebel (ms.), *AI and Consciousness*

Anil Seth (2026), “The Myth of Conscious AI”

Murray Shanahan (2024), “Simulacra as Conscious Exotica”

John Searle (1992), *The Rediscovery of the Mind*

Human likeness

"**[O]nly** of a living human being and what **resembles (behaves like)** a living human being can one say: it has sensations [...] is **conscious** or **unconscious**." (PI, 281)

- A **necessary**—not a sufficient—condition.
- Does not rule AI in or out.
- **Behavioral resemblance** will be crucial.