CAN COMPUTATION ESCAPE THE CHINESE ROOM?



Simulation, subjective experience, and the consciousness threshold

The Science of Consciousness 2026

Daniel Montoya | Fayetteville State University

If a system can flawlessly simulate understanding without ever "knowing" anything, is that consciousness—or just a Chinese Room with better wallpaper?

THE PROBLEM

Have you ever wondered why AI doesn’t incorporate Asimov’s 3 laws of Robotics?

1st Law

a robot may not harm a human or allow harm through inaction

2nd Law

A robot must obey human orders unless conflicting with the First Law (Second

3rd Law OF MIND

A robot must protect its own existence unless conflicting with the first two.

It’s not only because Tech Bros can’t read

AI can simulate understanding

LANGUAGE

“I understand what you mean.”

EMOTION

“That sounds painful. I’m sorry.”

THEORY OF MIND

“She believes he is mistaken.”

But does successful simulation imply subjective experience?

THREE LENSES

First Lens

Physical constraints

Second Lens

Computation and Semantics

Third Lens

Phenomenal Binding

Does the system has subjective experiences?

Could consciousness depend on more than computation?

COMPUTATION

What a system does
Rules, operations, transformations

PHYSICAL REALIZATION

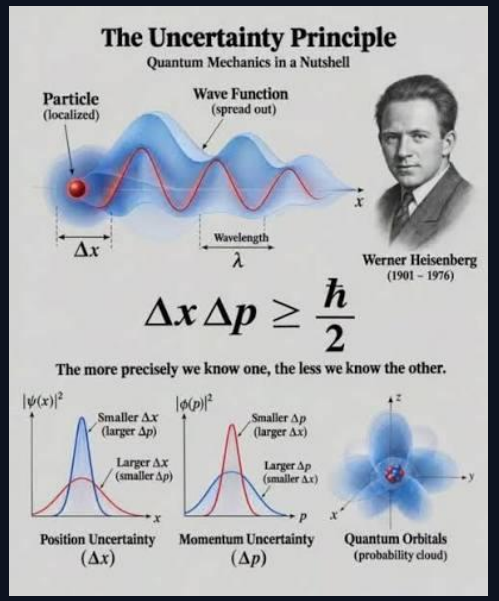
What physically carries it out
Neurons, silicon, living tissue

Every actual computation is physically realized.
The question is whether all physical realizations that implement the same computation are equivalent with respect to consciousness.

A possibility to explore

We do not know that consciousness requires non-classical physics.

If consciousness depends on physical properties beyond abstract computation, a digital simulation might reproduce the function without reproducing the phenomenon.



A possibility to explore

I.e.: Simulating pain is not necessarily feeling pain.

A computer can model the processes involved in pain and reproduce pain-related behavior. But if subjective experience depends on specific physical properties, reproducing the computation might not reproduce the experience.



Only one of these experience pain



The Chinese Room Argument (Searle, 1980)

A brief refresher.



Major Replies and Counter-Arguments

- **The Systems Reply**: The person inside the room does not understand Chinese, but the **entire system** (the person, the rulebook, the room, and the ledgers) collectively understands it. Comparing it to a computer, a single CPU core does not run the whole operating system by itself, but the integrated components produce functional cognition. Philosophy Stack Exch... +1
- **The Robot Reply**: A disembodied computer just shuffling text lacks interaction with the physical world. If you place the computer inside a **robot body** equipped with cameras and actuators, the machine can connect symbols to real-world objects, giving the symbols true semantic meaning. Stanford Encyclopedia ... +1
- **The Brain Simulator Reply**: If a program simulates the actual **neurobiological firing patterns** of a human brain processing Chinese step-by-step, that system replicates the causal properties of a real mind. Replicating the physical structure yields understanding. Stanford Encyclopedia ... +1
- **The Virtual Mind Reply**: A computer running a program can generate a **separate, distinct virtual agent** inside the architecture that understands the language, even if the host hardware or human operator remains oblivious to the meaning. Stanford Encyclopedia ... +1
- **The Other Minds / Intuition Reply**: Critics argue that Searle's conclusion relies heavily on **unreliable human intuitions** about what consciousness should look like. We attribute understanding to other human beings purely based on external behavior, so we should apply the same standard to machines. Stanford Encyclopedia ... +1

Bermudez, J.L. Cognitive Science: An Introduction to the Science of the Mind 4th Ed (2022)

The Chinese Room, updated

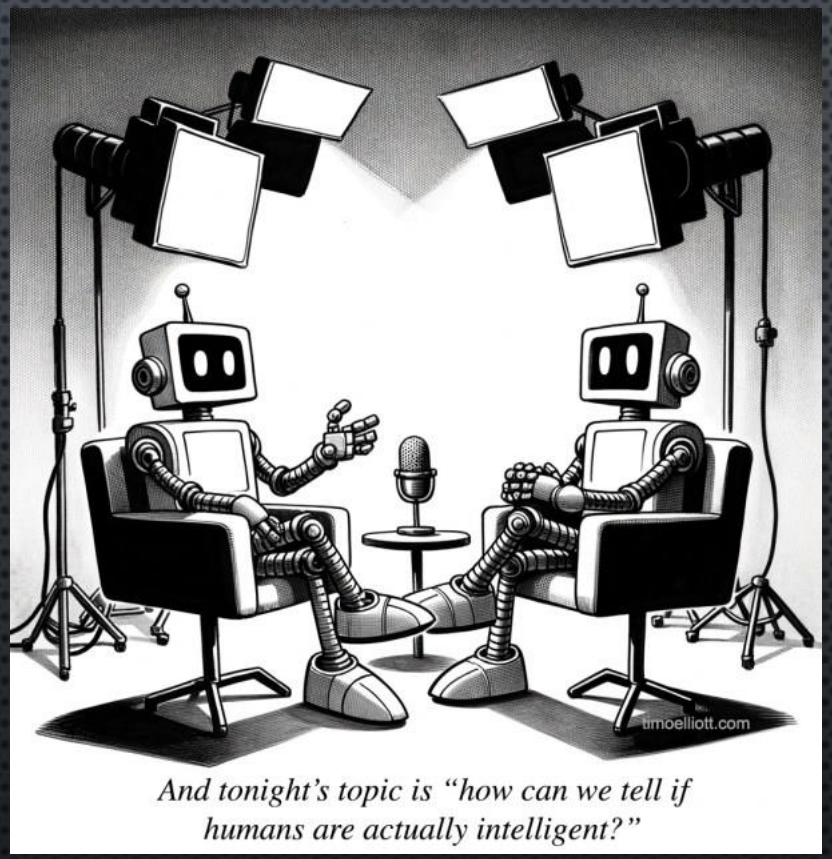


SYNTAX
Manipulating symbols



SEMANTICS
Understanding meaning

Does the right output prove genuine understanding?



And tonight's topic is "how can we tell if humans are actually intelligent?"

Simulated empathy is not necessarily felt empathy

AI RESPONSE

“I’m sorry you’re going through this.”

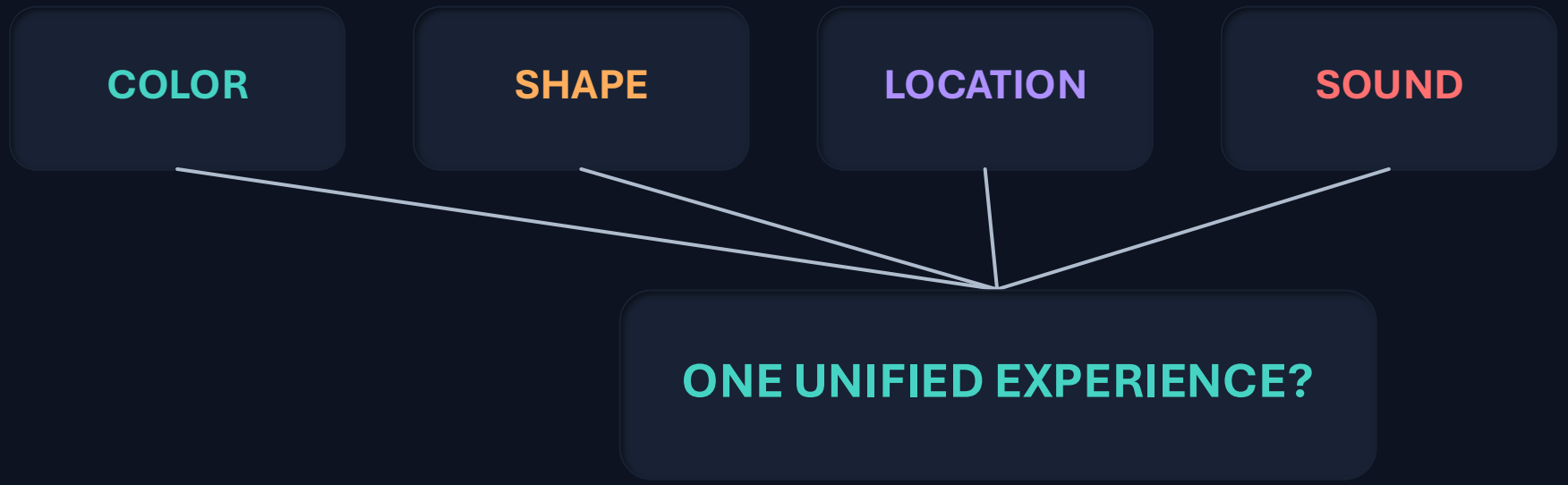
HUMAN EXPERIENCE?



The words may accompany felt concern.

Expression of emotion ≠ proof of emotional experience

How does “many” become “one”?





Different processes, yet experience seems unified.

Does discrete implementation preclude phenomenal unity?

YES

Discrete processing may be insufficient for phenomenal unity.



NO

Discrete components can produce unified functions.



UNKNOWN

Functional integration does not demonstrate experience.

A theoretical challenge, not an automatic disqualification of digital systems.

Functional binding ≠ phenomenal binding

A system can integrate information without showing why integration feels like anything

FUNCTIONAL BINDING

Feature A + Feature B
→ correct report/action

The system gets the relationship right.

PHENOMENAL BINDING

Feature A + Feature B
→ one experienced whole

Why should the integrated information constitute a unified experience?

* Percy & Argawal (2026) showed an ANN that can implement functional binding but fails to implement phenomenal binding

Percy & Agarwal, 2026, Consciousness and Cognition

All 3 lenses put together form a more detailed picture

Lens 1

Modeling pain is not the same as experiencing pain

Lens 2

A system can manipulate symbols according to rules and produce appropriate answers without that alone proving it understands their meaning

Lens 3

An ANN that can implement functional binding but fails to implement phenomenal binding

This is not an automatic disqualification of digital systems, but a theoretical challenge.

A POSSIBLE WAY FORWARD

What if consciousness is strongly emergent?



WEAK EMERGENCE

The whole arises from parts, while remaining explainable in principle from their interactions.

STRONG EMERGENCE

The whole has properties that cannot be fully derived from the properties of its parts.

Could consciousness be more than the sum of computational operations?

A hypothesis to explore.

A POSSIBLE WAY FORWARD

Why Strong Emergence Is often seen as better for explaining AI and Consciousness

If consciousness is only weakly emergent, then all mental states are ultimately reducible to computational or neural states.

In that case, a sufficiently detailed description of the system would completely explain consciousness. *



WEAK EMERGENCE

The whole arises from parts, while remaining explainable in principle from their interactions.

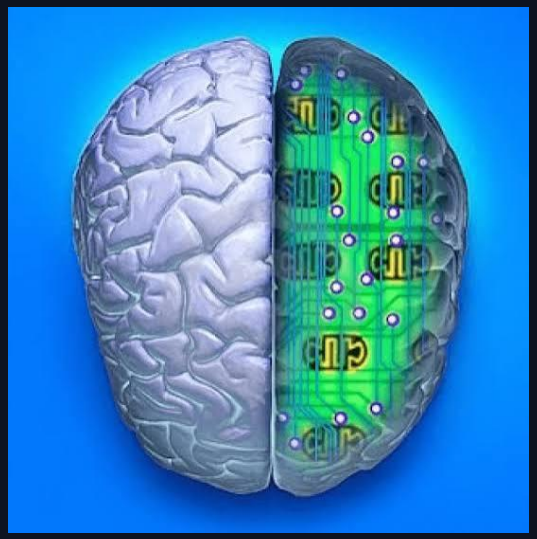
STRONG EMERGENCE

The whole has properties that cannot be fully derived from the properties of its parts.

*** Eerily similar to the Easy problem of Consciousness**

What would count as evidence?

FUNCTION	Can it behave intelligently?
SEMANTICS	Does it understand?
INTEGRATION	Is the information unified?
PHENOMENOLOGY	Is there something it is like to be the system?



What evidence would justify attributing subjective experience to a nonhuman system?

THE IDENTICAL-BEHAVIOR TEST

A thought experiment: can we distinguish simulation from awareness?

SYSTEM A | DIGITAL AI

Processes information digitally; recognizes emotions, solves problems, and reports an inner life (Introspection?).

SYSTEM B | BIOHYBRID

Uses living neural tissue connected to electronics; produces the same outward responses.

SAME OUTPUTS

Are either—or both—systems genuinely conscious, or are they simulating awareness?

Matching behavior alone may not settle whether subjective experience is present.

WHAT WOULD COUNT AS EVIDENCE?

Three research directions

1 FUNCTIONAL FLEXIBILITY

Can it adapt to unfamiliar situations and integrate conflicting information?

Shows more than rehearsed responses; does not prove experience. Tests for generalization.

2 CAUSAL INTEGRATION

Do perturbations to one process affect other processes and the system's coordinated response?

Tests how deeply information is integrated, not whether it is felt.

3 PHYSICAL DYNAMICS

Do different physical implementations produce differences beyond abstract computation?

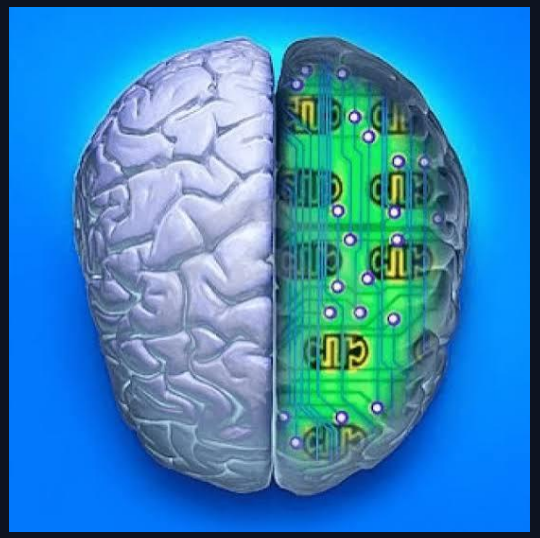
Could constrain theories; cannot by itself establish awareness.

Strong emergence / non-computable dynamics: hypotheses that must yield testable predictions.

Primary conclusions

Taken together, these considerations suggest that AI may approximate functional correlates of consciousness without ever instantiating consciousness itself.

Caveat: No single behavioral test can be assumed to settle the metaphysical question.



The biological wildcard

DIGITAL AI

Silicon-based computation



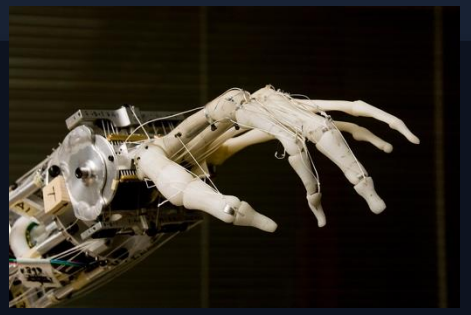
ORGANOID INTELLIGENCE

Living neural tissue in a controlled system



BIOHYBRID SYSTEMS

Biological and electronic components together



These technologies force us to ask what physical conditions consciousness actually requires.

The biological wildcard

DIGITAL AI

Silicon-based computation



ORGANOID INTELLIGENCE

Living neural tissue in a controlled system



Burst EEG activity

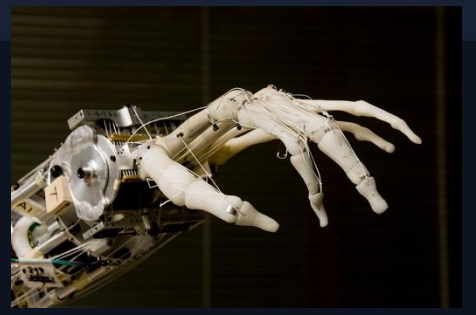
Show plasticity

Affected by psychedelics

Affected by anesthesia

BIOHYBRID SYSTEMS

Biological and electronic components together



These technologies force us to ask what physical conditions consciousness actually requires.

CONCLUSION

Three unresolved bridges

PHYSICAL

Could consciousness depend on properties beyond computation?

SEMANTIC

Does syntax ever become meaning?

PHENOMENAL

How does integration become experience?

Exploring whether strong emergence, potentially grounded in non-computable dynamics, offers a path forward, and propose criteria for distinguishing between sophisticated simulation and authentic awareness.

CONCLUSION

Three unresolved bridges

PHYSICAL

Could consciousness depend on properties beyond computation?

SEMANTIC

Does syntax ever become meaning?

PHENOMENAL

How does integration become experience?

Strong emergence may offer one possible bridge, but the question of subjective experience remains elusive

CAN COMPUTATION ESCAPE THE CHINESE ROOM?



Depends on how you ask

Is the simple accumulation of physical, semantic and phenomenal data enough to consider these artifacts conscious?

Depends on who you ask

Many seem already convinced that AI has reached consciousness at the price of ignoring the details of how AI achieves its results

CAN COMPUTATION ESCAPE THE CHINESE ROOM?

What would it take for a simulation to become an experience?

Thank you · Questions

SELECTED REFERENCES

Starting points for the argument

Chalmers, D. J. (1995). Facing up to the problem of consciousness. Journal of Consciousness Studies, 2(3), 200–219.

Dehaene, S., Charles, L., King, J.-R., & Marti, S. (2014). Toward a computational theory of conscious processing. Current Opinion in Neurobiology, 25, 76–84.

Kagan, B. J., et al. (2022). In vitro neurons learn and exhibit sentience when embodied in a simulated game-world. Neuron, 110, 3952–3969.e8.

Kosinski, M. (2024). Evaluating large language models in theory of mind tasks. PNAS, 121, e2405460121.

Revonsuo, A. (1999). Binding and the phenomenal unity of consciousness. Consciousness and Cognition, 8, 173–185.

Searle, J. R. (1980). Minds, brains, and programs. Behavioral and Brain Sciences, 3, 417–457.

Smirnova, L., et al. (2023). Organoid intelligence: The new frontier in biocomputing and intelligence-in-a-dish. Frontiers in Science, 1, 1017235.

Tononi, G. (2004). An information integration theory of consciousness. BMC Neuroscience, 5, 42.