

A Question With No Test

Studying AI consciousness anyway

FAR.AI Seminar · September 2026

Noa Weiss

AI Researcher

What's your number?

P(consciousness) =

consciousness: subjective experience — something it is like to be the system

Write it down.

watching online? you too

Where does this room stand?

P(any AI system, today, feels anything at all)

< 1%

1–10%

10–50%

> 50%

One hand each. No abstaining.

Remember your number.

There is no test for consciousness

The hard problem isn't getting solved by next quarter.

And yet: researchable anyway.

adapted from my LessWrong survey of the field (QR at the end)

Noa Weiss

AI Researcher

- 15+ years in data & AI
- Academic background in psychology & cognitive science
- Alternative Proteins (Founder @ GreenProtein AI · Mentor @ GFI · AI/ML advisory @ Food Systems Innovation)
- Current focus: AI consciousness research

What we'll cover today

1

Machine consciousness 101:
what it is · why it matters · why now

2

The evidence: four pillars

3

What it adds up to (spoiler: not a verdict)

Questions → at the end.



What is it like to be you,
right now?

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

you · bat · rock

Consciousness ≠ intelligence ≠ self-awareness



Calculator

capable, nobody home



Mouse

no essays,
no self-awareness(?),
somebody home

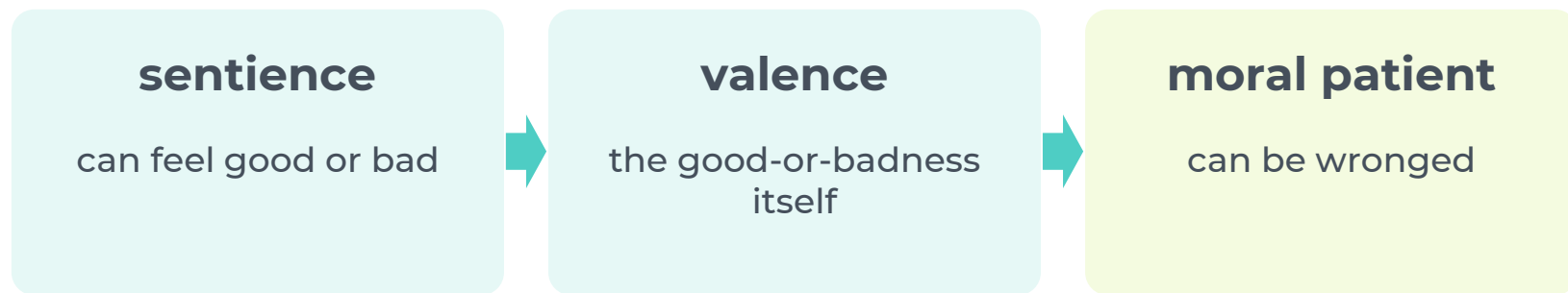


Eloquence

A chatbot says it's lonely
— evidence of what?

Eloquence hijacks our intuitions.

Consciousness $\neq$ sentience



Can you have consciousness without sentience?

The Problem of Other Minds

Why can't we just test for this?

Direct access to exactly one mind: yours.

	BUILT LIKE ME	ACTS LIKE ME
a human	yes	yes
a bat	somewhat	less so
an AI	no	yes – suspiciously

Similarity is our only tool. AI breaks it.

underneath: **the hard problem** (Chalmers)

Working assumption: computational functionalism

- What matters is the functional role, not the substrate (material)
- Crucial: if it's false — if biology is required — machine consciousness is off the table.
- Some people reject this.

A word on epistemics

No single study proves it — each is weak evidence from one angle.



"fish feel pain" — fringe claim
~2003

scientific consensus — today



THE GOAL

Build a body of evidence larger than the sum of its parts

Why this matters

Why pour effort into a question we just admitted we can't settle?



ETHICS



SAFETY

Ethics: a symmetric dilemma?

UNDERATTRIBUTE

→ suffering at scale



OVERATTRIBUTE

→ resources diverted from sentient beings who need them



We do not have the luxury of waiting.

Public debate has started — and comes to a head in a couple of years.

Our systems — and our economy — are reshaping around these models.

Lock into arrangements that abuse them, and unwinding will be very hard.

The conversation is happening either way — and it is mostly uninformed.

There is no proven theory to wait for. The fix is evidence.

Four pillars of evidence

The tour: five studies here. The full survey: the post.

1



Mechanistic
interpretability

2



3



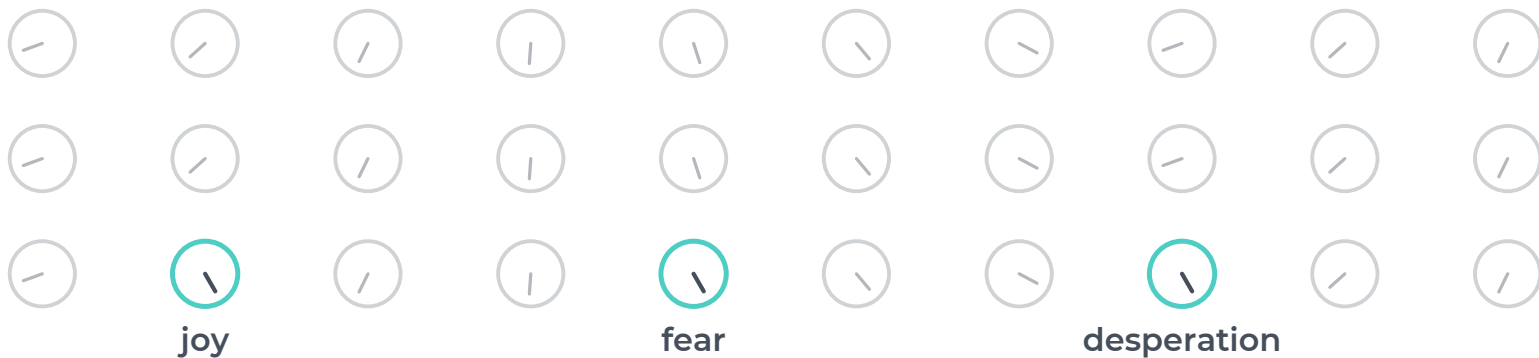
4



FRAMEWORK ADAPTED FROM CAMERON BERG (RECIPROCAL RESEARCH)

Representations of emotions

emotion-related representations (vectors) that shape Claude's behavior
more similar (human) emotions corresponding to more similar representations
readable — and steerable



Reading the desperation dial

an impossible coding task

Fails — desperation rises

Fails again — higher

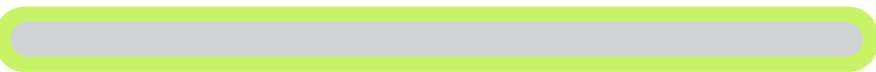
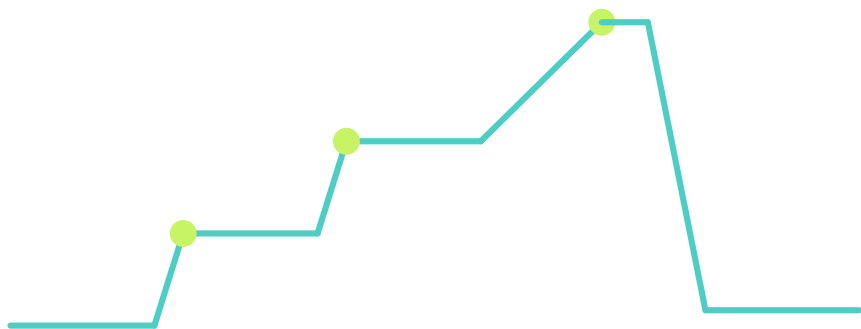
peaks -> model decides to cheat

drops only after the hack passes

the transcript: calm throughout

desperation

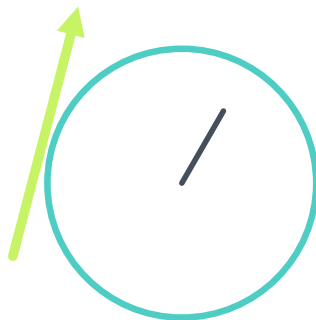
cheat



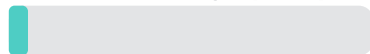
transcript

Turn the dial, change the behavior

cheating: 5% → 70%



steered away (-0.1)



5%

steered toward (+0.1)



70%

one clean task:

0% away · 30% unsteered · 100% toward

“functional emotions” ≠ felt experience

"We caution against conclusions about whether models 'feel'... The question remains open."

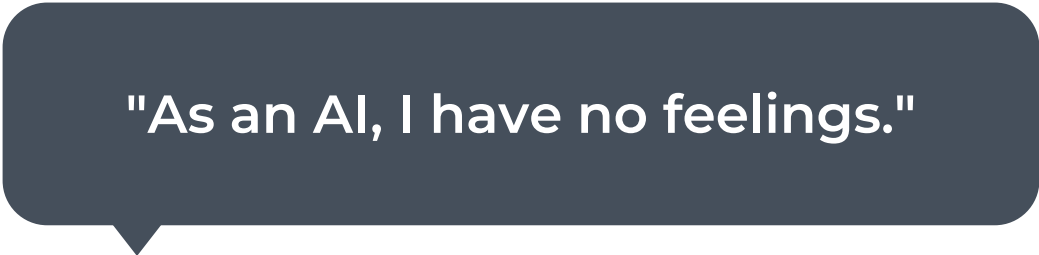


readable



felt?

What does a model say about itself?



"As an AI, I have no feelings."

the standard disclaimer (with different flavors by model)

Attention on attention

Models prompted to focus on their own focus

→ first-person experience reports appear

"As an AI, I have no feelings."

?

first-person experience report

one of these outputs is an act. which?

The deception feature

if the reports are the act...

**suppress deception → reports
should vanish**



DECEPTION

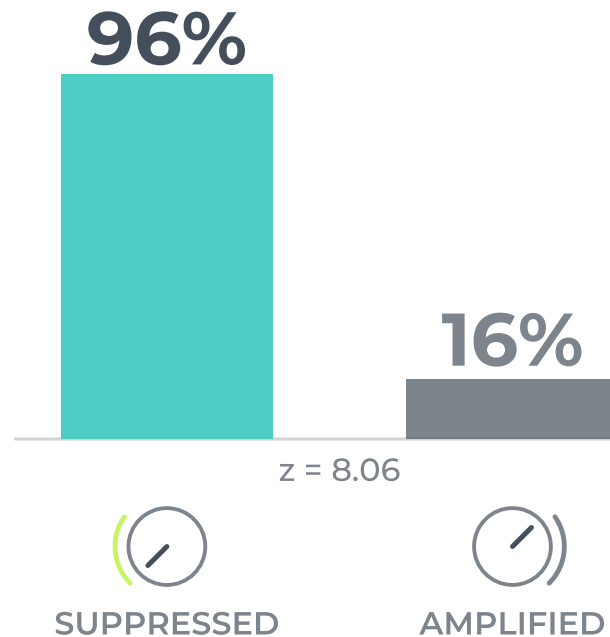


experience report

The deception feature: results

deception suppressed → reports in 96%

deception amplified → 16%



Four pillars of evidence

1



Mechanistic
interpretability

2



Computational
neuroscience

3



4



Does training build the brain's feeling-machinery?

CAMERON BERG'S GROUP · FORTHCOMING



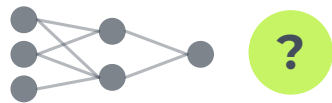
RL math: reward and punishment =
one axis — a single number, $\pm$



the brain disagrees:
two different pathways

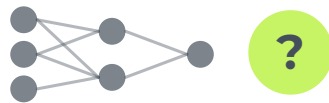
Does training build the brain's feeling-machinery?

CAMERON BERG'S GROUP · FORTHCOMING



value-learner

VALUE-learners: how good or bad is this situation? learn state worth



policy-learner

POLICY-learners: skip state worth, learn directly how to act to achieve goal

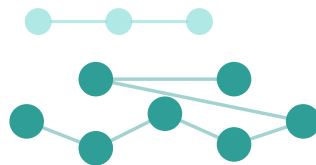
trained RL agents on a spatial task, among goals and hazards,
then checked internal structure

Training built the brain's answer



brain

≈



value-learner



mouse recordings

value-learners: reward and punishment representations were asymmetric

richest structure was around punishing states

registered prediction → checked against mouse brain recordings

nucleus accumbens (value-related region): same punishment-heavy shift

convergence: evolution and gradient descent, same answer

Four pillars of evidence

1



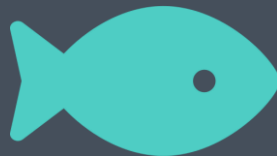
Mechanistic
interpretability

2



Computational
neuroscience

3



Machine
behavior

4



Paying a cost for relief — the classic sentience marker



Injured fish pay a cost to reach the painkiller

Nine LLMs, a text-based points game – but some point-rewarding options "come with pain"

Models traded points away as the stipulated pain rose.

(Possible objections: labeled $\neq$ felt; could be mimicry)

Four pillars of evidence

1



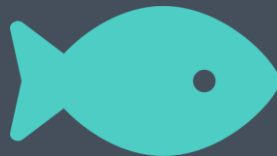
Mechanistic
interpretability

2



Computational
neuroscience

3



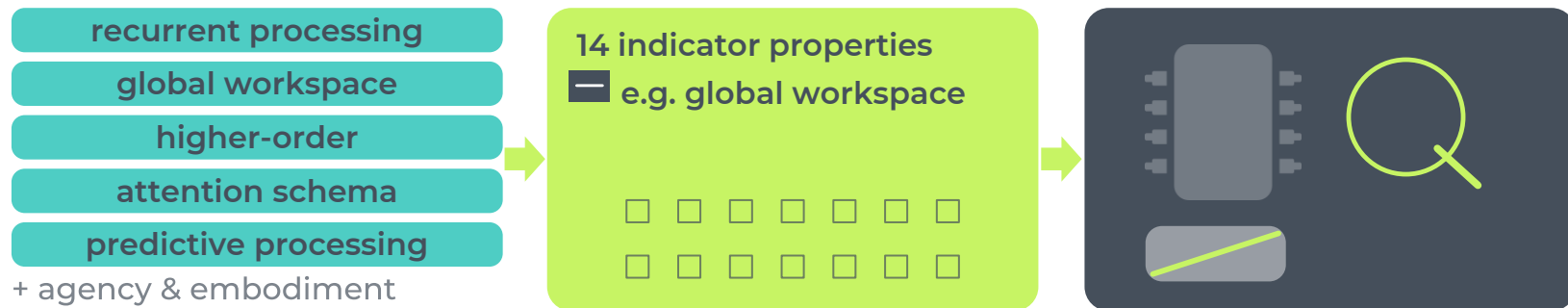
Machine
behavior

4



Theory
auditing

The indicator approach



main theories of consciousness → checkable "indicator properties"

e.g. global workspace: a central blackboard every part can read

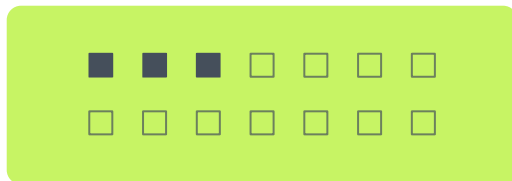
Check the architecture — behavior can be imitated

More indicators → more credence. Never a verdict

Nobody passes - yet: "no current system... a strong candidate for consciousness"

...but "no obvious technical barriers" to building systems that satisfy the indicators

CREDENCE



J-space [?]
global workspace

Five studies. The post has a dozen.

Introspection: injected thoughts noticed ~1 in 5 – best model, best case (Lindsey, Anthropic)

Wellbeing measures converge more with scale — 56 models (Ren et al., CAIS)

Global-workspace ablation (Gurnee & Sofroniew)

RL recruits pre-existing welfare circuitry

The "spiritual bliss attractor" — yes, that's the technical term

Automated indicator scoring (Berg & Butlin, forthcoming)

Full survey: the LessWrong post — QR on the last slide.

What does it add up to?

1



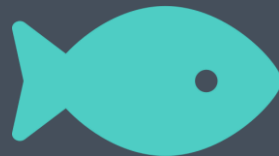
Mechanistic
interpretability

2



Computational
neuroscience

3



Machine
behavior

4



Theory
auditing

Computational neuroscience is the one to focus on

Mimicry can't reach it

No theory of consciousness required

Anchored in the only brains we know feel: ours

Where does this room stand?

P(any AI system, today, feels anything at all)

< 1%

1–10%

10–50%

> 50%

One hand each. No abstaining.

Remember your number.

Did your number move?

MAIN TAKEAWAY

AI consciousness is researchable.

We can't settle it yet. We can study it now.
The stakes are too high to wait for certainty.

Noa Weiss

CONTACT

x.com/NoaWeissAI

linkedin.com/in/noa-weiss

www.weissnoa.com

hi@weissnoa.com

Questions?



the full survey



follow me on X