The Case for Brain-Anchored Interpretability

ConCon · September 2026

Noa Weiss

AI Researcher

AI consciousness research: four ways, four holes

ASK IT

self-reports Perez & Long 2023

It read all we ever wrote about feeling; RLHF and instructions shape what it says about itself. Mimicry and training explain it all.

WATCH IT

behavior – the animal toolkit Keeling, Street et al. 2024

Same hole – it read about that too.

OPEN IT

interpretability Anthropic 2026 · functional emotions

Machinery says nothing about whether any of it is felt – and on its own, never can.

AUDIT IT

architecture vs. the theories Butlin, Long et al. 2023

Only as good as the theory you picked. The theories disagree.

THE FIFTH WAY

Brain-anchored interpretability

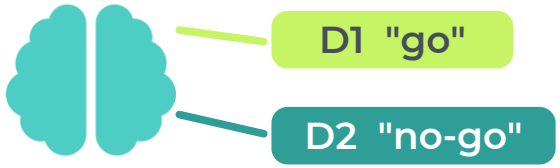
Mech interp, looking for mechanisms we already know from our own mammalian brains.

Rely on known theories from neuroscience and look for those in AIs.

One axis, or two?

ONE EXAMPLE · RECIPROCAL RESEARCH · FORTHCOMING

RL math: reward and punishment =
one axis – a single number, ±



the brain splits it: two pathways
(go / no-go – the dominant model)

Train agents among goals and hazards – look inside. **Results? Ask Cameron Berg**

Back to the holes

MIMICRY · INSTRUCTIONS

ask it · watch it

Neither can reach it – nothing here depends on what the model says or does.

WHICH THEORY?

audit it

No theory of consciousness to pick. The reference is a brain, not a theory.

REPRESENTED ≠ FELT

open it

Likeness to a system we know feels is the only bridge to 'felt' we have ever used. This builds it on purpose.

How do you know that I’m conscious?

WHY IT TRAVELS

The inference everyone already runs

Built like me → feels like I feel.

This is how humans already decide who is and isn't conscious.

The evidence that travels – into the public debate.

The evidence runs one way

WEAK EVIDENCE

If nothing brain-like turns up.

Machine consciousness might look nothing like ours – someone could be home, built differently.

STRONG EVIDENCE

If the brain's machinery turns up.

Trained networks have rediscovered brain structure before, unasked – convergence has precedent.

Noa Weiss

LET'S CONNECT

x.com/NoaWeissAI

linkedin.com/in/noa-weiss

www.weissnoa.com

hi@weissnoa.com

Agree? Disagree? Let's talk.

my survey of the field

follow me on X