The Case for Brain-Anchored Interpretability

Noa Weiss

ConCon · September 2026

There is still no test for consciousness, but there is one system we are sure has it – the brain – and we can look inside our models for the machinery it uses.

1 Every way we study AI consciousness has a hole in it

Way	Method	The hole	How brain-anchored interpretability (see 2) patches it
ASK IT	self-reports Perez & Long 2023	It has read everything we ever wrote about feeling, and training and instructions shape what it says about itself. Mimicry explains the data.	MIMICRY · INSTRUCTIONS Neither can reach this. Nothing here depends on what the model says or does.
WATCH IT	behaviour – the animal-welfare toolkit Keeling et al. 2024	Same hole: it has read about that too.	
OPEN IT	interpretability – functional emotions Sofroniew et al. 2026	On its own, machinery says nothing about whether any of it is felt. Nothing in a circuit diagram is labelled “felt”.	REPRESENTATION ISN’T FEELING Likeness to a system we know feels is the bridge to “felt” we have always used for other minds.
AUDIT IT	architecture vs. theories of consciousness Butlin et al. 2023, 2025	Only as good as the theory you picked. The theories disagree.	WHICH THEORY? None to pick. The reference is a brain.

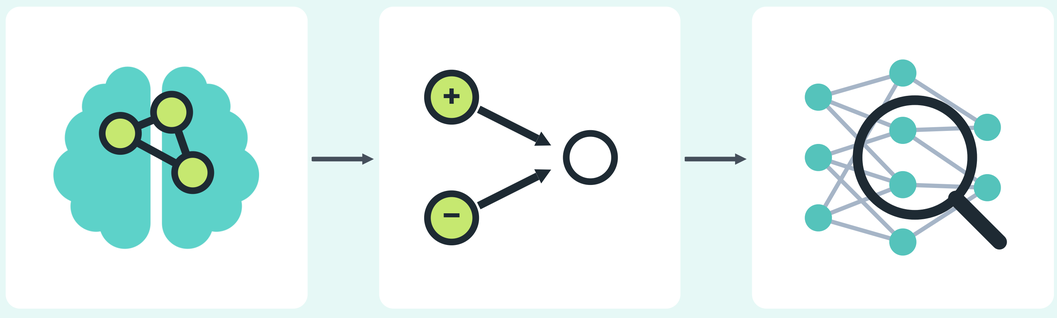
“Conscious” here means phenomenal: there is something it is like to be the system (after Nagel).

2 Brain-anchored interpretability

Open the model, as mechanistic interpretability does – but ask a different question. Not *what did training build*, but *did training build what the brain uses?*

Computational neuroscience has already mapped the circuits that implement affect in mammalian brains – the best-evidenced correlates of feeling we have. Look for those circuits inside the network. No scanner, no electrodes: neuroscience already did that work. The brain is the yardstick; the model is the specimen.

How it works



1. **Pick a mechanism.** A well-characterised circuit from the mammalian brain – one that ML theory alone would not predict a trained network to contain.
2. **Derive its signature.** Restate the circuit as a prediction in interpretability’s own vocabulary.
3. **Look inside – and decide what counts.** A match is the same computational structure, not the same substrate: same variables, same geometry, same causal role under intervention.

3 A caveat

Machine consciousness may not look like ours, which cuts both ways.

Nothing brain-like turns up: weak evidence. Someone could be home, built differently.

The brain’s machinery turns up: stronger evidence. And trained networks have rediscovered brain structure before, unasked (Yamins et al. 2014).

4 Why it travels

Humans already attribute consciousness by likeness: built like me, acts like me. This method uses only the first half – the half that cannot be mimicked.

“AIs have the brain’s feeling circuits” is a claim a non-specialist can weigh. The public debate is underway, and this is the evidence that will travel.

5 The ask

Mech interp researchers: you have read-write access. Neuroscientists: you have the maps. This work lives where the two meet – and too few people are building it.

My survey of the field

Me on X / Twitter

weissnoa.com
hi@weissnoa.com

Butlin, P., et al. (2023). Consciousness in artificial intelligence: insights from the science of consciousness. arXiv:2308.08708.

Butlin, P., et al. (2025). Identifying indicators of consciousness in AI systems. Trends Cogn. Sci. doi:10.1016/j.tics.2025.10.011.

Keeling, G., et al. (2024). Can LLMs make trade-offs involving stipulated pain and pleasure states? arXiv:2411.02432.

Nagel, T. (1974). What is it like to be a bat? Philos. Rev. 83(4).

Perez, E. & Long, R. (2023). Towards evaluating AI systems for moral status using self-reports. arXiv:2311.08576.

Sofroniew, N., et al. (2026). Emotion concepts and their function in a large language model. Transformer Circuits Thread.

Yamins, D. L. K., et al. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. PNAS 111(23).