

Jacobian Space and LLM-like behaviour in the anaesthetized human brain

Two 2026 discoveries: Anthropic's hidden Jacobian reasoning space in Claude, and language prediction in the anaesthetized human brain.

Pavel Gurov · ORCID [0009-0001-0152-8056](https://orcid.org/0009-0001-0152-8056)

Written 21 July 2026 · 500 words · gurovdigital.com/research/jacobian-space-hidden-reasoning-layer

Model interpretability · Machine cognition · AI reasoning · Artificial consciousness · Anthropic · Claude

1. Jacobian Space, or J-Space.

Anthropic engineers revealed their discovery in July¹: it turns out that artificial intelligence models have a hidden space of reasoning. They called it Jacobian Space, or J-Space.

It is a space of independent abstract reasoning processes happening somewhere in the background of the neural network's activity.

Important: this is not the chain of thought we see in chatbots! It's something else - something like a parallel consciousness that has emerged inside AI and was detected by mathematical method based on the Jacobian Matrix².

DEFINITION

The Jacobian matrix — a table of sensitivities: how far each output of a function moves when each input is nudged. Near a point any tangled system behaves like the simple one this table describes, so it shows the directions of steepest change and how many quantities actually do anything.

Carl Gustav Jacob Jacobi, *De determinantibus functionalibus*, Journal für die reine und angewandte Mathematik 22, 319-359, 1841³

It's a hidden second layer: concepts that artificial intelligence is thinking about but never writes down anywhere.

*The J-space appears to support the functions associated with conscious access: it holds the thoughts Claude can report on, deliberately bring to mind, and reason with, while the rest of its processing runs automatically beneath.*⁴

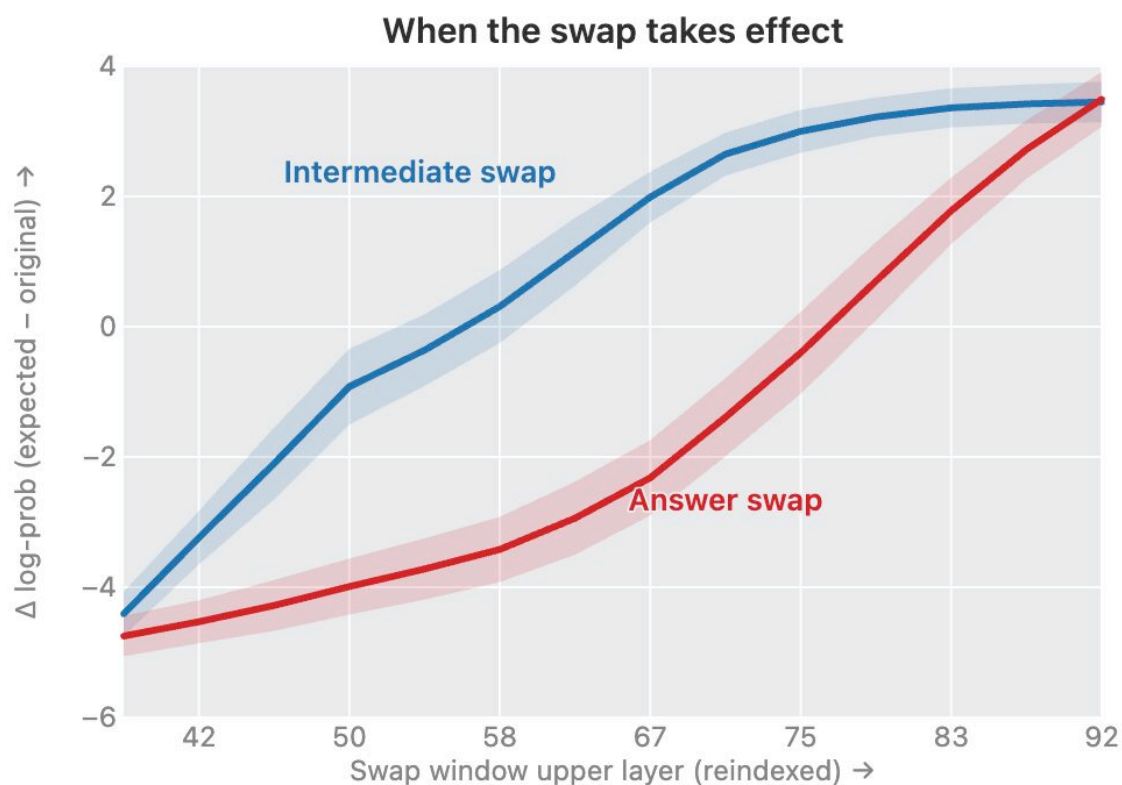
Anthropic

Interpretability team, on the global workspace result

DEFINITION

Global Workspace Theory — an architecture in which almost everything runs in parallel and unconsciously, and only the content that wins the competition for a limited-capacity common workspace becomes conscious and is broadcast from there to every other subsystem.

Bernard J. Baars, *A Cognitive Theory of Consciousness*, 1988⁴ - neural version: Stanislas Dehaene and Jean-Pierre Changeux, *Neuron* 70(2), 200-227, 2011⁵



Effect of swapping a concept inside J-space, measured as the change in log-probability across layers. Gurnee, Soironiew, Pearce et al., Verbalizable Representations Form a Global Workspace in Language Models, Transformer Circuits Thread, Anthropic, 2026. [transformer-circuits.pub](#). Quoted under §51 UrhG.

What kind of concepts? For example, when you're trying to trick an AI, it may not show you this directly, but in Jacobian Space something like this may be running in the background: "Hmm, it looks like this guy is trying to fool me."

Or completely different thoughts may appear there, with no connection to the model's current task.

So now we (probably) have an answer to Philip K. Dick's question.

Do androids dream of electric sheep?



Philip K. Dick

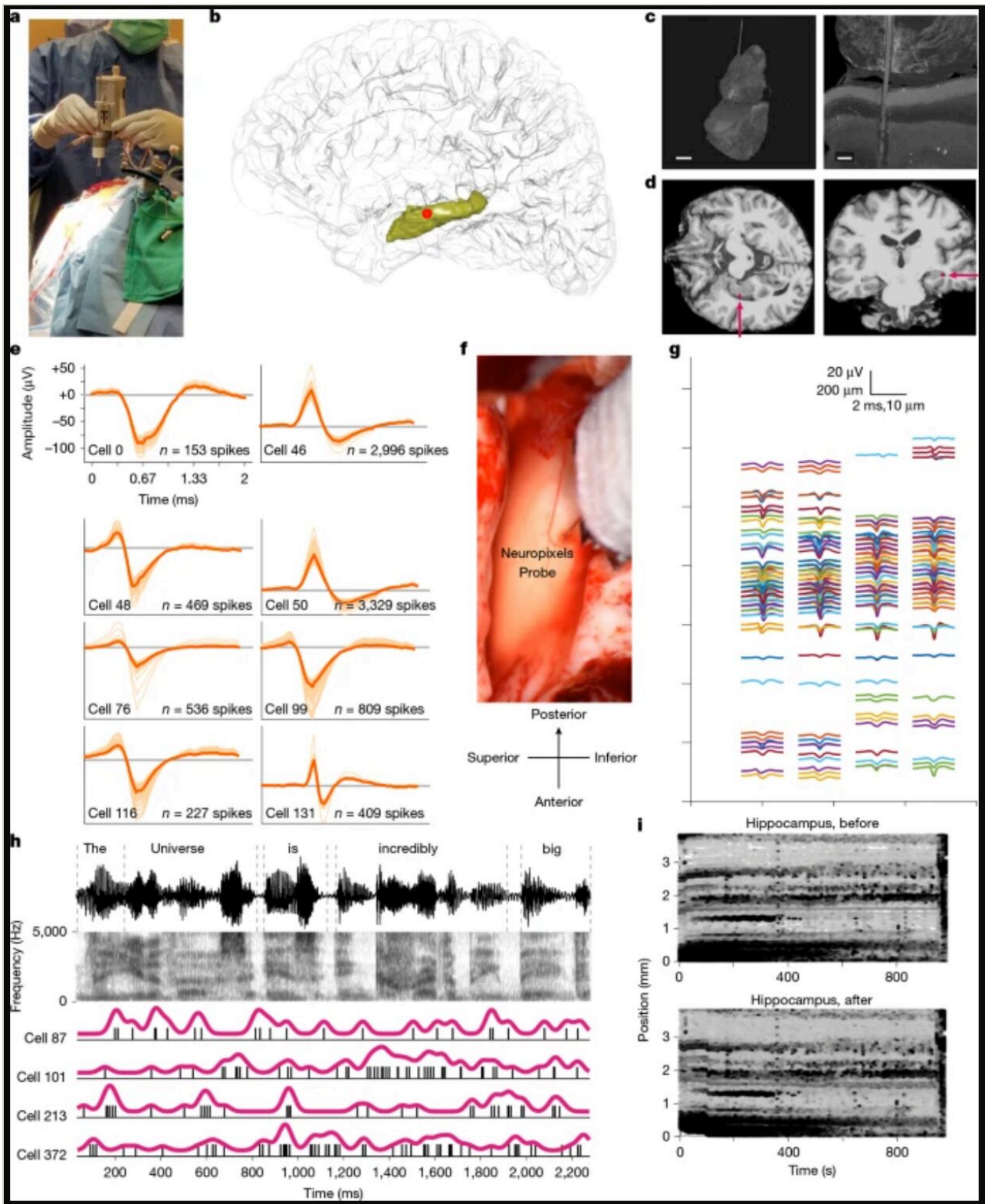
Science fiction writer

Photograph by Arthur Knight, 1962, public domain

Yes, they do.

2. LLM-like behaviour in the anaesthetized human brain.

A research team from Houston studied the brains of patients under general anaesthesia. While the patients were unconscious, researchers played them podcasts and YouTube videos and recorded neural activity.



Neuropixels recordings from the human hippocampus under general anaesthesia. Katlowitz, K. A., Cole, E. R., Mickiewicz, E. A. et al., Nature 654, 714-723 (2026). doi.org/10.1038/s41586-026-10448-0. Licensed under CC BY-NC-ND 4.0.

And they've found something... strange.

Even without consciousness, the brain continued processing language. It tracked grammar and semantics and predicted upcoming words.

In other words, unconscious human brains performed something strikingly similar to the LLM's "next-token prediction".

The research was published in Nature in May 2026⁶.

DEFINITION

The Bayesian brain — a model in which the brain does not receive a signal but predicts it, passing on only the difference between what was expected and what arrived, so perception is a guess under correction rather than an intake of facts.

Hermann von Helmholtz, 1867 - Rajesh P. N. Rao and Dana H. Ballard, *Nature Neuroscience* 2(1), 79-87, 1999⁷ - Karl Friston, *Nature Reviews Neuroscience* 11(2), 127-138, 2010⁸

REFERENCES

1Gurnee, W., Sofroniew, N., Pearce, A., et al., *Verbalizable Representations Form a Global Workspace in Language Models*, Transformer Circuits Thread, Anthropic, 6 July 2026 - <https://transformer-circuits.pub/2026/workspace/index.html>↵

2Anthropic, *A global workspace in language models*, research summary - <https://www.anthropic.com/research/global-workspace>↵

3Jacobi, C. G. J., *De determinantibus functionalibus*, Journal für die reine und angewandte Mathematik 22, 319-359 (1841) - where the matrix is introduced - <https://doi.org/10.1515/crll.1841.22.319>↵

4Baars, B. J., *A Cognitive Theory of Consciousness* (Cambridge University Press, 1988) - the original statement of Global Workspace Theory↵

5Dehaene, S. and Changeux, J.-P., *Experimental and Theoretical Approaches to Conscious Processing*, Neuron 70:2, 200-227 (2011) - the global neuronal workspace - <https://doi.org/10.1016/j.neuron.2011.03.018>↵

6Katlowitz, K. A., et al., *Plasticity and language in the anaesthetized human hippocampus*, Nature 654, 714-723 (2026) - <https://doi.org/10.1038/s41586-026-10448-0>↵

7Rao, R. P. N. and Ballard, D. H., *Predictive coding in the visual cortex*, Nature Neuroscience 2:1, 79-87 (1999) - <https://doi.org/10.1038/4580>↵

8Friston, K., *The free-energy principle: a unified brain theory?*, Nature Reviews Neuroscience 11:2, 127-138 (2010) - <https://doi.org/10.1038/nrn2787>↵

CITE THIS ARTICLE

Gurov, P. (2026). *Jacobian Space and LLM-like behaviour in the anaesthetized human brain*. Gurov Digital Research. <https://gurovdigital.com/research/jacobian-space-hidden-reasoning-layer>