The Internal Dynamics of Decisions: From rats and primates to AI

Brain, or AI model? *Tell them apart by their dynamics.* · **Nathan M. Thornhill**

Institute for Complexity Science and Advanced Computing (ICSAC), Fort Wayne, Indiana · nathanthornhill.com · nthornhill@icsacinstitute.org

POSTER #530

tsc2026.nathanthornhill.com/preprint

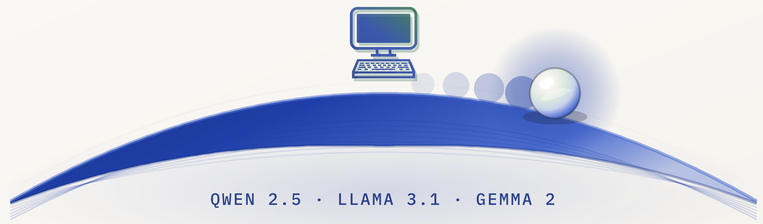


ANIMAL BRAIN CORTEX · RUNS ON ITSELF

Can you tell them apart?

tsc2026.nathanthornhill.com

Interactive game
Preprint
References
Pre-registrations
Data
Bio



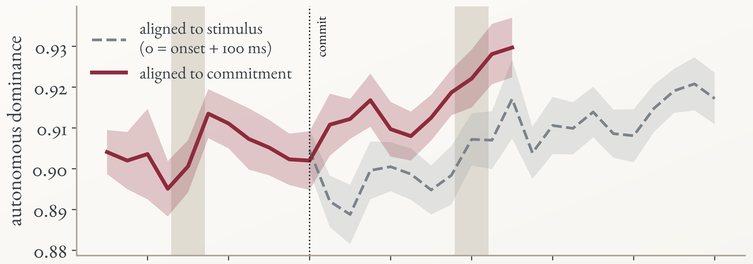
AI MODEL · RUNS ON ITS INPUT

BRAIN

The rat decides. Then its cortex holds on longer.

75 of 115

SESSIONS HELD ON LONGER AFTER COMMITMENT

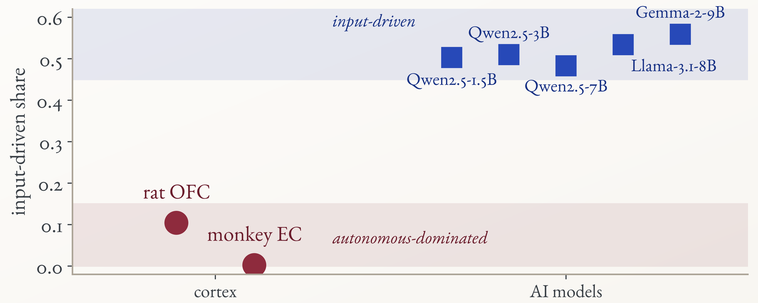


How strongly cortex holds its own state, before and after the rat commits. Shaded: the measurement windows. 12 rats · about 36,000 units. Aligned to the clicks instead (dashed), the rise is smaller and more gradual. Specificity to commitment is supported, not established.

Under harder trials rat orbitofrontal cortex leans further on itself. The monkey recordings lean the same way, in 10 of their 14 sessions.

Attribution. Luo et al. reported this transition in 2025, and the method's authors ran their tool on the same recordings. It's recovered here, not discovered. What's new is putting a brain and an AI model on the same axis.

Brains run on themselves | AI models run on what you feed them



How much of each next moment arrives from outside, rather than being generated by the system itself. Plotted as first analyzed, at unmatched input dimension: read regime, not magnitude.

A feedforward network can copy a brain's behavior, so behavior alone can't separate the two. Whatever separates them is inside.

The two brain recordings are a rat working out how long a reward is worth waiting for and a monkey tracking a path it can no longer see, in different parts of the brain on jobs that have nothing to do with each other. Both sit on the same side.

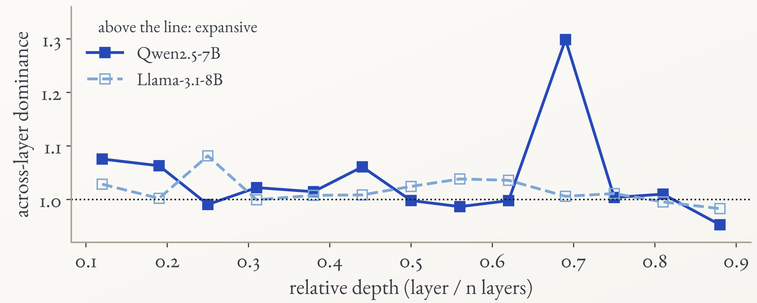
Read the sides, not the sizes. Re-run at a matched input dimension (rat OFC, two Qwen models), the sides held. But the AI models' label depends on how their input is counted, so the distances don't compare.

AI MODEL

The answer surfaces near the last layer. Then it runs out of layers.

0 of 54

TOKEN-LEVEL TESTS FOUND A SIGNIFICANT RISE AFTER COMMITMENT



Layer by layer the read-out hovers near or above 1. The tall spikes are output-projection artifacts. A forward pass is a feedforward chain, so this read-out is not evidence either way.

In both models tested, the answer first becomes top-ranked, under the readout used, only near the **last layer** (28 of 29 · 32 of 33). The rat's cortex holds on longer *after* the choice. Within a forward pass there is no post-commitment interval to measure, so the test ran across generated tokens instead: two pre-registered passes, 30 tests in five AI models, then 24 in four. Window and criterion were fixed in advance, so a hit would have counted.

The measure separates input from autonomy

N4SID, a subspace state-space fit. It fits the population as

`x(t+1) = Ax(t) + Bu(t)`

A is what the population does to itself. **Bu** is what the input does to it. The measure is how much of each next step comes from **A**.

What a population does next depends on what you fed it, and what you fed it depends on what it just did. Ordinary methods hand back a blend, and on these recordings they find **nothing at all**.

Told to find **0.60** in synthetic systems with a known answer, it never missed by more than 0.002, across a fourfold sweep of input strength.

Borrowed, not built here. The estimator is N4SID (Van Overschee & De Moor 1994). The input-aware demix is InputDSA (Huang et al. 2025), on Ostrow et al. 2023.

Wiring is the only difference

The looping ones carry the signature: 3 of 3 on the delayed task.

Six small networks built from scratch, not the five AI models above. **Three loop** (RNN, GRU, LSTM), **three don't** (MLP, TCN, causal transformer). All six trained on **the click-counting task the rats did** until they **all scored the same**, and matched **within 1.13×** on parameter count. So it isn't that one performs better, and it isn't that one is bigger.

144 networks, none thrown out, and the whole design was **frozen before** a single dynamical measure was computed.

Four runs in all, two versions of the task crossed with two independent seed sets. All four came out in the predicted direction, including both on seeds it had never seen, the weakest at about 1 in 80 by chance and the strongest better than 1 in 100,000. Counted by architecture instead of by network, all three looping designs sit above all three others on the delayed task.

The nulls that keep it honest

Pre-registered tests in rat OFC, read trial-averaged or as one scalar per trial, found no commitment signal. Committed against abandoned choices: nothing. Opting in against opting out: nothing. A scalar read-out across 9,106 trials: nothing. OFC is a value area, not the commitment circuit, so these bound the result, and a measure that returns something under every analysis isn't measuring anything.

The other pre-registered measure doesn't separate the six networks. Only the eigenvalue measure does. Both were specified in advance and both are reported.

The monkey data is thin. Two animals, 14 sessions. It agrees with the rat on the axis and under harder trials, but it can't carry a result by itself, and the rise on the left was measured in rats.

The demix cannot certify that cortical autonomy is recurrent rather than functionally feedforward (Goldman 2009). The architecture claim rests on the six-network control, where the wiring is known.

Why this matters

For thirty years one question has been whether access needs **recurrent** dynamics or whether **feedforward** processing could do the same job. The unfolding argument says behavior can never tell you, because a feedforward network can copy any recurrent brain's behavior. That gets treated as the end of the road.

It isn't the end of the road. It means the answer is inside, and inside can be measured. Wire six networks two ways, train them to the same score on the same task, and their dynamics still tell them apart. **"It behaves the same" was never the same as "it works the same".**

None of this says a machine is or isn't conscious. Recurrent processing theory (Lamme 2006), global workspace ignition (Mashour et al. 2020), integrated information theory (Tononi et al. 2016) all tie access to internal dynamics. A rise in autonomous persistence at commitment is one property of that kind. It is in cortex and was not found in the five AI models tested. Whether that difference matters for access is for those theories to say.

Data: Schiereck et al. 2026 · Neupane et al. 2024 · Luo et al. 2025 · Method: Huang et al. 2025 · Ostrow et al. 2023 · Van Overschee & De Moor 1994 · Full record, references and both pre-registrations behind either code