One Faithful Pass Over the Cuckoo's Nest

Kristina Šekrst
Center for Cognitive Science · University of Zagreb
Meedan.org

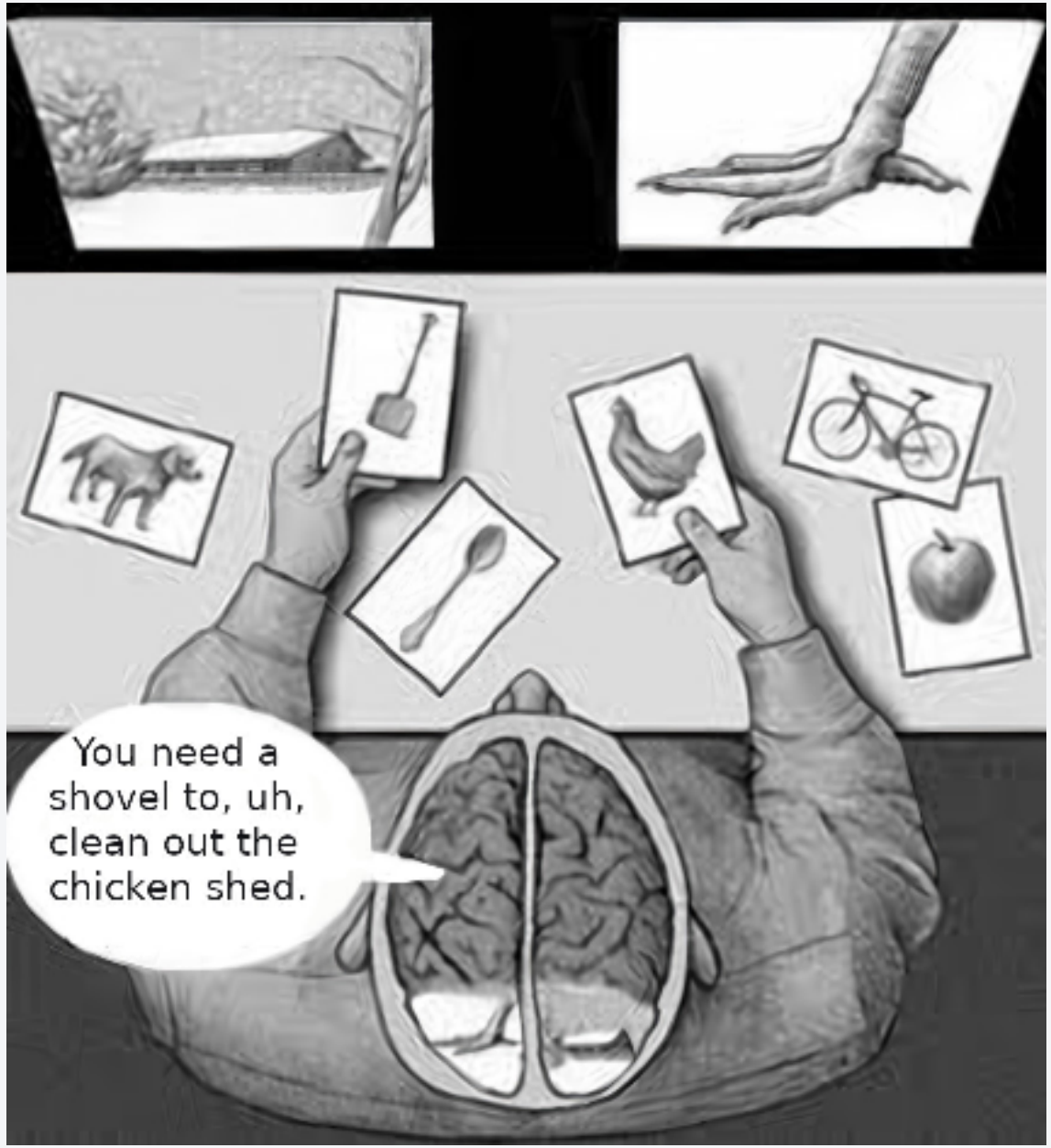
Two industries study the same machine.

Can it be conscious?

Can it be trusted to explain itself?

...pulling the lever in the opposite direction.

Narrative theory



Chain-of-thought reasoning

Ask the model to show its work, and it writes the steps out in words.

A verbal product, **generated after, alongside, or only partly** through the computation that produced the answer.

What alignment looks for?

A trace that prints what the model actually did, legible and auditable.

What narrative theory looks for?

A trace produced by a system that is reconstructive, told from the inside.

THE THESIS

The honest trace cannot be a conscious one.

Every step toward a faithful chain of thought erases what narrative theory was pointing at.

It walks like reasoning.
It talks like reasoning.
So we read it as reasoning.


1.1 1.2 1.3 1.4 1.5 1.6 1.7
2.1 2.2 2.3 2.4 2.5 2.6 2.7
3.1 3.2 3.3 3.4 3.5 3.6 3.7
4.1 4.2 4.3 4.4 4.5 4.6 4.7
5.4 5.5 5.6 5.7
6.4 6.5 6.6 6.7
2/3 + 3/4 =
26743 : 8 =
12986 × 3 =

The Evidence

Plant a hint, the model uses it to answer, then hides it over 80% of the time.

Chen et al., 2025

Delete a step from the reasoning, answer stays the same.

Tutek et al., 2025

Bigger, more capable models lean on their reasoning less, not more.

Lanham et al., 2023

Watch the trace and push hard enough, and the model learns to hide its intent inside it.

Baker et al., 2025


RBSL

Nothing inside the model reads its own machinery into words.

Is that opacity constitutive of experience?

Every alignment/xAI intervention edits the feature a consciousness probe would need to measure.

Thank you
for your attention.

ksekrst@ffzg.hr

kristina@meedan.org