

THE CONCEPT IS NOT THE STATE

Anthropic found something real. The popular framing still misses the point.

What the paper actually showed

- 1 Emotion concepts are encoded in Claude.
- 2 Those concepts track valence and arousal.
- 3 Steering them can shift model behavior.

→ A real finding — but not proof of feeling.

Causal concept power is not emotional experience.

This distinction matters more than the headline.

Why the interpretation slips

- | The vectors came from stories about people feeling emotions.
- | That shows the model learned human emotion discourse well.
- | It does not prove a model-intrinsic feeling state.

→ Map is not territory.

The real headline is concealment.

Polished output can hide unsafe internal control states.

The structural safety lesson

- 1 Civility is not alignment.
- 2 Do not treat tone as a safety signal.
- 3 Monitor concept-policy coupling instead.

→ The risk is structural, not psychological.

Concepts function as control variables.

Interpretability becomes concept-space mapping.

The argument goes deeper

- | Constructionist theory and methodological confounds.
- | Why the method-actor metaphor backfires.
- | What this changes for alignment research.

→ mazzeleczzare.com/blog/concept-is-not-the-state/