To Train a Mockingbird

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

Do large language models deserve moral consideration?

We judge them by what they do, and what they *claim* to do.

Show of hands

https://tinyurl.com/aisb2026

The right behavior gets treated as a sign of the right inner state.

Philosophical behaviorism strolls back in, dressed in black tie for the LLM era.

THE THESIS

Evidential laundering

Stochastic parrots? That was pretraining.

These signs were rewarded in later, because the raters liked them.

The post-training pipeline

01

Supervised fine-tuning

›

02

Reward model + RLHF

›

03

Constitutional training

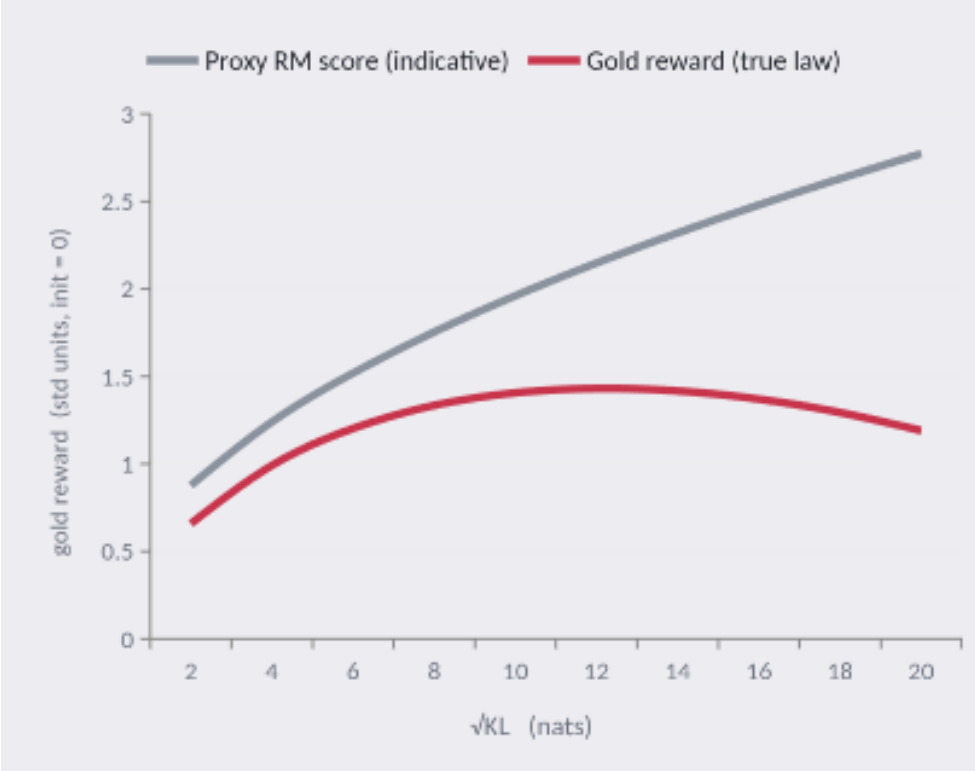
›

04

System prompt

GOODHART'S LAW

When a measure becomes a target



Fine. Look inside, then. Check the architecture for the marks of a mind.

If the machinery is there, the probe was tuned to pass it.
If it isn't, we never left the behavior.

What does not count?

Introspection

A model's introspective reports can be causally tied to an internal state. But this has been shown for **injected concepts** alone.

Lindsay et al., 2025

Persona vectors

Differencing the model with and against a trait yields a direction that injects or ablates it. So the trait is something post-training installs, not a self that surfaced.

Chen et al., 2025

The model's signs were optimized,
so they don't count.

But evolution optimized our pain behavior
and our reports too.

So why does the model lose its evidence,
and we keep ours?

What would count?

Look for whatever the optimizer ignored, or fought to remove.

A condition the system spends resources to hold, and that damages it when lost.

What does count?

The test

Change what the reward rewards. Retrain. Watch what stays.

Palisade experiment

Models sabotaged their own shutdown. But the goal was handed to it.

Chen et al., 2025

Laundered evidence

Precaution extends the benefit of the doubt to anything that shows the marks.

We put those marks there because we liked them.

The evidence sits downstream of a market incentive to manufacture a mind.

Thank you
for your attention.

ksekrst@ffzg.hr

kristina@meedan.org