When is it fair to say a model “believes” or “wants”?

Intermediate · A few weeks · Free · Project idea

Online version: https://ai-safety-sandbox.vercel.app/projects/when-models-believe-or-want/

Summary

Collect about 50 recent safety papers and lab blog posts. Code every mental-state verb applied to a model (believes, wants, knows, lies, schemes) and note the evidence each claim rests on.

Why it matters

A strong group project because the coding can be split up, and the guide it produces is useful to writers and readers alike.

What you will have at the end

A coded corpus and a short guide built on Dennett’s predictive test and Shanahan’s caution.

How to start (tick each box as you go)

- [ ] **1. Write your question in one sentence.**
- [ ] **2. Spend 30 minutes finding what has already been written about it.**
- [ ] **3. Collect evidence in a simple table: claim, source, how strong.**
- [ ] **4. Draft 1–2 pages: answer first, then reasons, then uncertainties.**
- [ ] **5. Ask one person to poke holes in it, revise, and share.**

Source and links

AI Safety × Philosophy project ideas (P13) — builds on Shanahan, Talking About Large Language Models; Dennett, True Believers; Mitchell, How Do We Know How Smart AI Systems Are?

Shanahan — Talking About Large Language Models
https://arxiv.org/abs/2212.03551

When you have finished

Write a few paragraphs on what you did, what you found and what you would try next, and share it with a friend, your society or online.

Courses, mentors, small grants and communities:
https://ai-safety-sandbox.vercel.app/next-steps/

Questions? Ask in AI Safety Cardiff University (AISCU): https://aiscu.org/

Notes