Spot the mind-words in AI research posts

Starter · 1–3 hours · Free · Step-by-step guide

Online version: https://ai-safety-sandbox.vercel.app/projects/believes-wants-audit-pilot/

Summary

Read three AI lab blog posts and highlight every time a model is said to “believe”, “want”, “know”, “lie” or “scheme”. For each, ask: what evidence backs this claim up?

Why it matters

Mental-state language shapes how the public and policymakers think about AI. It is sometimes justified and sometimes not, and noticing the difference is a skill.

What you will have at the end

A short coded list and a paragraph of guidance.

Steps (tick each box as you go)

- [ ] **1. Pick three posts (15 min)**
  Choose recent research posts from AI labs or safety organisations.
- [ ] **2. Highlight the verbs (30 min)**
  Mark every mental-state word applied to a model.
- [ ] **3. Code the evidence (30 min)**
  For each, note what the claim rests on: behaviour, internal evidence (interpretability), or nothing in particular.
- [ ] **4. Draft a rule of thumb (15 min)**
  Using Dennett’s idea that belief-talk is justified when it helps predict behaviour, write one sentence on when such language seems fair.

What you need

- A web browser
- About 90 minutes

What you will learn

- The intentional stance
- Reading research critically
- How language shapes AI debates

Source and links

Short version of a project from the AI Safety × Philosophy reading list (P13)

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.

Try next: When is it fair to say a model “believes” or “wants”?
https://ai-safety-sandbox.vercel.app/projects/when-models-believe-or-want/

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