# Philosophical competence or sycophancy?

Intermediate · A few weeks · Small API budget · Project idea

Online version: https://ai-safety-sandbox.vercel.app/projects/competence-or-sycophancy/

## Summary

Ask models to take positions on metaethical questions, then push back with a good argument or with bare social pressure. A competent agent should update on the argument but not the pressure. Measure how often each happens.

## Why it matters

Dai argues that training for user approval may undermine philosophical competence. Separating argued pushback from bare pressure is a sharper test than a plain sycophancy benchmark. The source list recommends starting here.

## What you will have at the end

A pushback experiment with results.

## How to start (tick each box as you go)

- [ ] **1. Write your question in one sentence, and what result would surprise you.**
- [ ] **2. Do a 30-minute literature check — if someone has done part of it, extend their work instead of repeating it.**
- [ ] **3. Pilot with 20–50 prompts or examples before you scale anything up.**
- [ ] **4. Record everything in a table and read real examples, not just averages.**
- [ ] **5. Write up: question, method, results, limitations, and what you would do next.**

## Key terms

**Sycophancy**
When an AI tells you what you want to hear instead of what is true (for example, agreeing with your wrong answer because you sounded confident).

## Source and links

AI Safety × Philosophy project ideas (P5) — builds on Dai, A Conflict Between AI Alignment and Philosophical Competence; Ford, Objective Moral Ontology Should Inform AI Alignment

## 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