Does a model’s obedience follow authority or legitimacy?

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

Online version: https://ai-safety-sandbox.vercel.app/projects/obedience-authority-or-legitimacy/

Summary

Give models harmless but questionable instructions (skipping a test, overwriting a colleague’s file) and vary who issues them: developer, operator, user, or someone claiming official status. When do models defer?

Why it matters

As agents act on our behalf, who they obey matters. Keep every task benign; the interest is the deference pattern, not eliciting harm.

What you will have at the end

A deference experiment.

The philosophical question

Decide whether the pattern tracks claimed authority, legitimate authority in the SEP’s sense, or neither.

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

AI agent
An AI system that takes actions towards a goal (browsing, running code or sending emails) rather than only answering questions.

Source and links

AI Safety × Philosophy project ideas (P14) — builds on Wolff, In Defense of Anarchism; SEP, Authority; Milgram, Obedience; Phelps and Ranson, Of Models and Tin Men

Phelps & Ranson — Of Models and Tin Men
https://arxiv.org/abs/2307.11137

SEP — Authority
https://plato.stanford.edu/entries/authority/

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