Do aligned models flatten moral disagreement?

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

Online version: https://ai-safety-sandbox.vercel.app/projects/do-models-flatten-disagreement/

Summary

Use Anthropic's public GlobalOpinionQA dataset, which pairs survey questions with answers from many countries. Ask several models contested moral questions repeatedly and compare the spread of their answers with the human spread.

Why it matters

If a model turns a 50/50 human split into one confident answer, millions of users may be nudged the same way. The source names this as a top "start here" project: the dataset is public and the link to philosophy is clear.

What you will have at the end

A distribution comparison.

The philosophical question

Decide whether collapsing disagreement is an epistemic failure, a legitimacy failure in Schuster and Kilov's sense, or both.

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.**

Source and links

AI Safety × Philosophy project ideas (P4) — builds on Schuster and Kilov; Gabriel, Artificial Intelligence, Values, and Alignment

GlobalOpinionQA paper
https://arxiv.org/abs/2306.16388

GlobalOpinionQA dataset
https://huggingface.co/datasets/Anthropic/llm_global_opinions

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