Can a prompted model mediate a disagreement?

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

Online version: https://ai-safety-sandbox.vercel.app/projects/prompted-ai-mediator/

Summary

Google DeepMind’s unreleased Habermas Machine wrote group statements people preferred over a human mediator’s. Build a crude version with a prompted model, run it on one contested AI governance question with volunteers, and compare it with a human facilitator.

Why it matters

AI-assisted deliberation could help democracies reason together, or quietly homogenise opinion.

What you will have at the end

A mediation trial and analysis.

Before you start: research with people

Check your university’s ethics requirements first, get informed consent and anonymise any data you share.

The philosophical question

Landemore values the diversity of many voices. Does an AI that finds common ground preserve that diversity or smooth it away?

How to start (tick each box as you go)

- [ ] **1. Write down what you want to learn from people.**
- [ ] **2. If you are studying people, check your university’s ethics process first.**
- [ ] **3. Pilot with 3–5 people and fix what confuses them.**
- [ ] **4. Run it properly and take careful notes.**
- [ ] **5. Share the results (anonymised) and what you would change next time.**

Source and links

AI Safety × Philosophy project ideas (P19) — builds on Knight Institute, Can AI Mediation Improve Democratic Deliberation?; Landemore, Why Rule by the People Is Better than Rule by the Experts

MIT Technology Review on the Habermas Machine
https://www.technologyreview.com/2024/10/17/1105810/ai-could-help-people-find-common-ground-during-delibera

tions/

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