Run your own collective constitution

Starter · A few weeks · Free · Project idea

Online version: https://ai-safety-sandbox.vercel.app/projects/student-collective-constitution/

Summary

In Anthropic and CIP's study, about 1,000 Americans used the open-source tool Polis to write a constitution for an AI. Run the same process with 30–100 students and compare your result with theirs.

Why it matters

It needs no compute, works as a society event and produces something to compare with a published study. The source list recommends starting here.

What you will have at the end

A Polis process, a constitution, and a comparison.

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

Does a student group's constitution have any legitimacy? Use the SEP's instrumental and intrinsic justifications of democracy to argue what a process like this can and cannot confer.

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 (P18) — builds on Anthropic and CIP, Collective Constitutional AI; SEP, Democracy

Collective Constitutional AI
https://www.anthropic.com/research/collective-constitutional-ai-aligning-a-language-model-with-public-input

Polis
https://pol.is/home

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