What values does AI safety research encode?

Intermediate · A few weeks · Free · Project idea

Online version: https://ai-safety-sandbox.vercel.app/projects/values-encoded-in-safety-research/

Summary

Apply Birhane et al.'s annotation scheme to about 50 recent alignment and safety papers, and compare the values they emphasise with the ML papers in the original study.

Why it matters

Does the safety field's stated aim show up in what its papers actually value? This works well as a group project, with the annotation split between members.

What you will have at the end

An annotated paper set and short report.

How to start (tick each box as you go)

- [ ] **1. Write your question in one sentence.**
- [ ] **2. Spend 30 minutes finding what has already been written about it.**
- [ ] **3. Collect evidence in a simple table: claim, source, how strong.**
- [ ] **4. Draft 1–2 pages: answer first, then reasons, then uncertainties.**
- [ ] **5. Ask one person to poke holes in it, revise, and share.**

Source and links

AI Safety × Philosophy project ideas (P6) — builds on Birhane et al., The Values Encoded in Machine Learning Research

Birhane et al. — The Values Encoded in Machine Learning Research
https://arxiv.org/abs/2106.15590

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