Map the tensions between corrigibility and a constitution

Intermediate · A few weeks · Free · Project idea

Online version: https://ai-safety-sandbox.vercel.app/projects/corrigibility-vs-constitution/

Summary

List every place a published AI constitution asks a model to act on its own values, and every place it asks the model to accept correction. Find where both apply and conflict.

Why it matters

Real AI developers have to resolve this tension in writing. Mapping it clearly helps everyone see the choices being made.

What you will have at the end

An argument map that resolves each conflict using Wolff's account of autonomy and Christiano's definition of corrigibility, and says which resolution the documents actually choose.

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.**

Key terms

Corrigibility
An AI's willingness to be corrected, modified or shut down by the people responsible for it, rather than resisting.

Source and links

AI Safety × Philosophy project ideas (P15) — builds on Christiano, Corrigibility; Google DeepMind, An Approach to Technical AGI Safety and Security (§6.5.1); Anthropic, Claude's Constitution

Anthropic — Claude's Constitution
https://www.anthropic.com/constitution

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