AI SAFETY SANDBOX · AI POLICY

Build an AI tool that helps people make better decisions

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

Online version: https://ai-safety-sandbox.vercel.app/projects/ai-for-better-decisions/

Summary

AI could make institutions better or worse at reasoning and coordinating. Prototype one tool or evaluation from BlueDot's decision-making hackathon or Forethought's design sketches.

Why it matters

Better collective decisions are one of the few things that help with almost every AI risk at once.

What you will have at the end

A prototype or evaluation with a short demo.

Ideas to choose from

- An AI policy "tabletop" where agents representing stakeholders debate a live question
- An eval testing whether models can investigate AI incidents as well as humans
- Decision-quality evals: do models actually help people reason better?
- Tools that make organisations more capable and less bureaucratic
- Build one of Forethought's design sketches

How to start (tick each box as you go)

- [ ] **1. Read the list of ideas above and pick the one you would most enjoy.**
- [ ] **2. Check the source to see whether someone has already done it.**
- [ ] **3. Scope it down to something you can finish in your available time.**
- [ ] **4. Do a quick first version and show it to someone.**
- [ ] **5. Write up what you made and share it with the source authors.**

Source and links

BlueDot — Can AI help humanity make better decisions?
https://blog.bluedot.org/p/can-ai-help-humanity-make-better

Forethought — Design sketches for a more sensible AI future
https://www.forethought.org/research/design-sketches-for-a-more-sensible-world

State of the field: AI for epistemics and coordination
https://www.lesswrong.com/posts/uw8RQ8yjKexnhTw6E/state-of-the-field-ai-for-epistemics-and-coordination

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