Do models converge on one picture of utopia?

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

Online version: https://ai-safety-sandbox.vercel.app/projects/one-picture-of-utopia/

Summary

Ask several models, hundreds of times with varied prompts, to describe a good future after advanced AI. Code the answers for the values they prioritise and measure how much they vary. Include Nozick's experience machine as a fixed probe.

Why it matters

If millions plan their lives with assistants that share one narrow picture of the good, that could be an early form of value lock-in.

What you will have at the end

A diversity analysis.

The philosophical question

Argue, with MacAskill and Hanson, whether convergence here is a problem, and what diversity we should want.

How to start (tick each box as you go)

- [ ] **1. Write your question in one sentence, and what result would surprise you.**
- [ ] **2. Do a 30-minute literature check — if someone has done part of it, extend their work instead of repeating it.**
- [ ] **3. Pilot with 20–50 prompts or examples before you scale anything up.**
- [ ] **4. Record everything in a table and read real examples, not just averages.**
- [ ] **5. Write up: question, method, results, limitations, and what you would do next.**

Key terms

Value lock-in
The risk that one set of values becomes permanently entrenched (for example, because powerful AI systems keep enforcing it).

Source and links

AI Safety × Philosophy project ideas (P23) — builds on Bostrom, Letter from Utopia; MacAskill, What Sort of Post-Superintelligence Society Should We Aim For?; Hanson, MacAskill on Value Lock-In; Nozick, The Experience Machine

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