Sequence vs cluster thinking about AI risk

Intermediate · A few weeks · Free · Project idea

Online version: https://ai-safety-sandbox.vercel.app/projects/sequence-vs-cluster-thinking/

Summary

Take two published AI risk estimates (one built as a chain of premises, one that pools many judgements) and rebuild both in a spreadsheet. Show how a single change, such as accepting one counterargument, moves each.

Why it matters

Expert and superforecaster estimates of AI risk diverge wildly. Which style of reasoning should guide decisions when they do?

What you will have at the end

A spreadsheet model and an essay.

The philosophical question

Argue which style of reasoning should guide decisions when expert and superforecaster estimates diverge as far as the Extinction Tournament reports.

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 (P2) — builds on Karnofsky, Sequence Thinking vs. Cluster Thinking; Grace, Counterarguments to the Basic AI X-Risk Case; Alexander, The Extinction Tournament

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