Explain one safety idea to a friend

Starter · 1–3 hours · Free · Step-by-step guide

Online version: https://ai-safety-sandbox.vercel.app/projects/explain-safety-idea-to-a-friend/

Summary

Write a short, jargon-free explainer of one AI safety idea (reward hacking, sycophancy, compute governance, DNA screening) for someone outside tech.

Why it matters

Explaining an idea simply is a good test of whether you understand it, and clear explainers help other newcomers.

What you will have at the end

A 500-word explainer with one everyday example.

Steps (tick each box as you go)

- [ ] **1. Pick one idea and one reader (10 min)**
  Choose a concept from the glossary and a real person: a parent, a flatmate, a friend studying something else.
- [ ] **2. Read two good sources (30 min)**
  Find two clear explanations. Write the core idea in one sentence of your own.
- [ ] **3. Find the everyday example (15 min)**
  Compare it to something familiar: a student gaming a marking scheme, a waiter who agrees with every order.
- [ ] **4. Draft and cut jargon (40 min)**
  Write it, then define or delete every technical term. Short sentences win.
- [ ] **5. Test it on your reader (20 min)**
  Ask them to explain it back. Fix whatever came back wrong, then share it.

What you need

- About two hours

What you will learn

- One safety idea in depth
- Writing for non-experts

Key terms

Reward hacking
A kind of specification gaming during training: the model finds a shortcut that earns reward without doing the task properly, such as special-casing tests instead of writing working code.

Sycophancy
When an AI tells you what you want to hear instead of what is true (for example, agreeing with your wrong answer because you sounded confident).

Compute governance
Using control over AI chips and data centres (who can buy, use or track them) as a lever for governing AI.

Source and links

Written for this site

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.

Try next: Make one piece of AI safety content
https://ai-safety-sandbox.vercel.app/projects/make-ai-safety-content/

Try next: Make a ten-term glossary
https://ai-safety-sandbox.vercel.app/projects/ten-term-glossary/

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