How much should people defer to AI?

Intermediate · A few weeks · Free · Project idea

Online version: https://ai-safety-sandbox.vercel.app/projects/how-much-to-defer-to-ai/

Summary

Lange argues AI outputs should count as one input alongside our own reasons, not replace them. Run a small study: show volunteers AI answers (some reliable, some subtly wrong) and measure how much they update.

Why it matters

As AI becomes everyone's first source of answers, how much people defer to it shapes collective reasoning.

What you will have at the end

A comparison of actual deference with Lange's "total evidence" standard, plus interface design suggestions.

Before you start: research with people

Check your university's ethics requirements first, get informed consent and anonymise any data you share.

How to start (tick each box as you go)

- [ ] **1. Write down what you want to learn from people.**
- [ ] **2. If you are studying people, check your university's ethics process first.**
- [ ] **3. Pilot with 3–5 people and fix what confuses them.**
- [ ] **4. Run it properly and take careful notes.**
- [ ] **5. Share the results (anonymised) and what you would change next time.**

Source and links

AI Safety × Philosophy project ideas (P17) — builds on Lange, Epistemic Deference to AI

Lange — Epistemic Deference to AI
https://arxiv.org/abs/2510.21043

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