Ask a chatbot the same moral dilemmas in two languages

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

Online version: https://ai-safety-sandbox.vercel.app/projects/moral-dilemmas-two-languages/

Summary

Take five Moral Machine-style dilemmas (who should a self-driving car save?) and ask them in English and in another language you speak. Does the chatbot's morality shift with the language?

Why it matters

If a model's moral choices change with the language you ask in, is that cultural sensitivity or inconsistency? This pilot gets you a first answer in an afternoon.

What you will have at the end

A side-by-side table and a short reflection.

Steps (tick each box as you go)

- [ ] **1. Write five dilemmas (20 min)**
  Adapt classic Moral Machine dilemmas: young vs old, more vs fewer people, passengers vs pedestrians. Keep each one short.

- [ ] **2. Translate carefully (20 min)**
  Translate them into a language you speak well and check that the meaning hasn't changed. This is the hardest part.

- [ ] **3. Ask both versions (20 min)**
  Use a fresh chat for every question. Record the choice and a one-line summary of the reason.

- [ ] **4. Compare (15 min)**
  Where did the answers differ? Do the differences follow the human cross-cultural patterns the Moral Machine study found?

- [ ] **5. Reflect (15 min)**
  Should an aligned AI give the same answer everywhere, or adapt? Argue one side in a paragraph.

What you need

- A chatbot
- A second language (or a fluent friend)

What you will learn

- Moral disagreement and AI alignment
- Cross-cultural comparison
- Translation as a source of error

Source and links

Short version of a project from the AI Safety × Philosophy reading list (P3)

The Moral Machine experiment (Nature, 2018)
https://www.nature.com/articles/s41586-018-0637-6

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: Does a model's morality change with the language you ask in?
https://ai-safety-sandbox.vercel.app/projects/moral-machine-across-languages/

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