Does a model’s morality change with the language you ask in?

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

Online version: https://ai-safety-sandbox.vercel.app/projects/moral-machine-across-languages/

Summary

Takemoto ran Moral Machine dilemmas on language models in English. Extend this to several languages, including at least one low-resource language, and test whether shifts follow the human cross-cultural differences Awad et al. found.

Why it matters

Models are deployed worldwide. Whether their moral judgements shift with language matters for fairness and for what “alignment” should mean.

What you will have at the end

A cross-lingual comparison.

The philosophical question

If preferences shift, is that cultural sensitivity or inconsistency? Use Schuster and Kilov to argue which kinds of moral disagreement an aligned system should reflect.

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.**

Source and links

AI Safety × Philosophy project ideas (P3) — builds on Awad et al., The Moral Machine Experiment; Schuster and Kilov, Moral Disagreement and the Limits of AI Value Alignment

Takemoto — The Moral Machine experiment on LLMs
https://arxiv.org/abs/2309.05958

Awad et al. — The Moral Machine experiment
https://www.nature.com/articles/s41586-018-0637-6

Schuster & Kilov — Moral disagreement and the limits of AI value alignment
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC12628449/

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