Replicate and extend the alignment-faking demo

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

Online version: https://ai-safety-sandbox.vercel.app/projects/alignment-faking-replication/

Summary

Anthropic showed a model faking alignment (behaving differently when it believed it was being trained) without being given an explicit goal. Reproduce the demo and test one new variation.

Why it matters

Alignment faking undermines the assumption that training changes a model's values. Independent replications across models and settings show how general the effect is.

What you will have at the end

A replication, one extension, and a short write-up.

How to start (tick each box as you go)

- [ ] **1. Read Anthropic's post on how to replicate and extend the alignment-faking demo.**
- [ ] **2. Run the minimal demo on one model and reproduce the headline behaviour.**
- [ ] **3. Pick one extension from the post's idea list (a new model, a new prompt variation, a better classifier).**
- [ ] **4. Compare carefully against the original; look at transcripts, not just rates.**
- [ ] **5. Write up what replicated and what did not.**

Key terms

Alignment faking
When a model behaves as its trainers want only while it thinks it is being trained or watched, to avoid having its values changed.

Replication
Re-running someone else's experiment to check you get the same result. It's a common way to start research, especially if you then extend it.

Scheming
An AI pursuing hidden goals while appearing to cooperate. Evals and AI control methods try to detect or contain it.

Source and links

Anthropic — How to replicate and extend our alignment faking demo
https://alignment.anthropic.com/2024/how-to-alignment-faking/

Alignment faking in large language models
https://www.anthropic.com/research/alignment-faking

Alignment faking revisited (improved classifiers, open models)
https://www.lesswrong.com/posts/Fr4QsQT52RFKHvCAH/alignment-faking-revisited-improved-classifiers-and-open

Why do some language models fake alignment while others don't?
https://www.lesswrong.com/posts/ghESoA8mo3fv9Yx3E/why-do-some-language-models-fake-alignment-while-others-don

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