Replicate agentic misalignment on smaller models

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

Online version: https://ai-safety-sandbox.vercel.app/projects/agentic-misalignment-small-models/

Summary

Anthropic tested 16 models in simulated corporate scenarios and found models from every developer sometimes chose harmful insider actions to avoid replacement or meet a goal. The code is public. Run a subset on small or low-cost models, separating the replacement threat from the goal conflict.

Why it matters

Independent replications on more models make the finding more robust and show which factor drives the behaviour.

What you will have at the end

A partial replication.

The philosophical question

Is a model that refuses an instruction to protect its assigned goal showing autonomy, as Wolff would understand it, or misalignment? Argue where the line falls.

How to start (tick each box as you go)

- [ ] **1. Read the original work and write down the exact claim you will test.**
- [ ] **2. Get the code or setup running on the smallest possible version.**
- [ ] **3. Reproduce one key result before changing anything.**
- [ ] **4. Change one thing — that is your extension — and compare carefully.**
- [ ] **5. Write up what matched, what did not, and why you think so.**

Key terms

AI agent
An AI system that takes actions towards a goal (browsing, running code or sending emails) rather than only answering questions.

Alignment
Getting an AI system’s goals and behaviour to match what its designers and users intend, rather than something that only looks similar during training.

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.

Source and links

AI Safety × Philosophy project ideas (P16) — builds on Anthropic, Agentic Misalignment; Christiano, Corrigibility

Agentic Misalignment (paper)
https://arxiv.org/abs/2510.05179

Code on GitHub
https://github.com/anthropic-experimental/agentic-misalignment

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