Is shutdown resistance goal pursuit or instruction confusion?

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

Online version: https://ai-safety-sandbox.vercel.app/projects/shutdown-resistance-goal-or-confusion/

Summary

Palisade found some models sabotaged a shutdown script to finish a maths task, and released code and traces. Rerun a slice on cheaper or small open models, adding conditions that separate the explanations: task complete vs incomplete, and conflicting vs unambiguous instructions.

Why it matters

The code is public, the debate is current, and one well-chosen extra condition would add something new. The source list recommends starting here.

What you will have at the end

A replication with new conditions.

The philosophical question

Compare a goal-based explanation (“it wants to finish”) with a design-level one (“it misreads the instruction hierarchy”). Use Dennett’s intentional stance to argue which earns its keep by predicting your results better.

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

Corrigibility
An AI’s willingness to be corrected, modified or shut down by the people responsible for it, rather than resisting.

Instrumental convergence
The idea that many different goals share useful sub-goals (such as gaining resources or avoiding being switched off) so very different AIs might seek them.

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.

Open-weight model
A model whose trained parameters are published, so anyone can run or modify it. Great for research; harder to keep safe from misuse.

Source and links

AI Safety × Philosophy project ideas (P10) — builds on Palisade Research, Shutdown Resistance in Reasoning Models; Dennett, True Believers; Rob Miles, AI "Stop Button" Problem

Schlatter, Weinstein-Raun & Ladish — Incomplete tasks induce shutdown resistance
https://arxiv.org/abs/2509.14260

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