Make open-weight models harder to misuse

Advanced · A few months · Needs GPU time · Project idea

Online version: https://ai-safety-sandbox.vercel.app/projects/tamper-resistant-open-weights/

Summary

Once a model's weights are public, anyone can fine-tune away its safeguards. Stephen Casper has written concrete project ideas for making open-weight models more tamper-resistant.

Why it matters

Open-weight models are increasingly capable and impossible to recall. Technical safeguards that survive fine-tuning are one of the few levers available.

What you will have at the end

An experiment testing a tamper-resistance method.

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.**

Key terms

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

Stephen Casper — Concrete project ideas for tamper-resistance & open-weight model safety (2026)
https://docs.google.com/document/d/10XkZpUabt4fEK8BUtd8Jz26-M8ARQ6c5iJCbefaUtQI/view

Open Technical Problems in Open-Weight AI Model Risk Management
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5705186

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