Run an AI control experiment with ControlArena

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

Online version: https://ai-safety-sandbox.vercel.app/projects/try-ai-control-setting/

Summary

AI control asks: if a model might be trying to sabotage us, can we still use it safely? ControlArena is the UK AISI's open toolkit of settings, model organisms and protocols. Run one setting, then improve a monitor or document a gap.

Why it matters

Control is one of the most practical near-term safety agendas, and it rewards ordinary software and security skills.

What you will have at the end

A documented experiment or a contribution to ControlArena.

How to start (tick each box as you go)

- [ ] **1. Read the ControlArena README and run its example setting.**
- [ ] **2. Write down the "red team" (attacker) and "blue team" (monitor) strategies in plain English.**
- [ ] **3. Change one thing (a monitor prompt, a threshold, an attack) and measure the effect.**
- [ ] **4. Open an issue or PR for anything confusing you hit along the way.**
- [ ] **5. Write up your result in a short post.**

Key terms

AI control
Designing safeguards that keep AI systems from causing harm even if they might be misaligned (for example, monitoring a model with another model).

Model organism
A deliberately built example of a worrying behaviour in a model, made so it can be studied safely (like a lab mouse for AI risks).

Red-teaming
Deliberately trying to make a system fail or misbehave, so weaknesses are found and fixed before bad actors find them.

AI Safety / Security Institute
A government body that tests advanced AI models and researches their risks, such as the UK's AI Security Institute.

Source and links

BlueDot — The software engineer's guide to your first AI safety contribution
https://blog.bluedot.org/p/swe-ai-safety-project-guide

ControlArena on GitHub
https://github.com/UKGovernmentBEIS/control-arena

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