Pick an open problem in technical AI governance

Intermediate · A few months · Free · Project idea

Online version: https://ai-safety-sandbox.vercel.app/projects/open-problem-in-technical-ai-governance/

Summary

Reuel et al. map open problems by capacity (assessment, access, verification, security, operationalisation and ecosystem monitoring) across data, compute, models and deployment. Choose one cell and scope a project.

Why it matters

It is the most cited map of where technical people can help governance. Starting from it means your project plugs into a real need.

What you will have at the end

A scoped project plan, then a first result.

How to start (tick each box as you go)

- [ ] **1. Write your question in one sentence.**
- [ ] **2. Spend 30 minutes finding what has already been written about it.**
- [ ] **3. Collect evidence in a simple table: claim, source, how strong.**
- [ ] **4. Draft 1–2 pages: answer first, then reasons, then uncertainties.**
- [ ] **5. Ask one person to poke holes in it, revise, and share.**

Key terms

Verification
Ways to check, rather than just trust, that someone is following an agreement (for example, confirming a data centre is not secretly training a banned model).

Evaluation (eval)
A repeatable test of what an AI model can do or how it behaves. Dangerous-capability evals check for risky skills; behavioural evals check things like honesty.

Source and links

Reuel et al. — Open Problems in Technical AI Governance (2024)
https://arxiv.org/abs/2407.14981

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