Explain one way to verify an AI agreement

Intermediate · A few weeks · Free · Project idea

Online version: https://ai-safety-sandbox.vercel.app/projects/explain-a-verification-mechanism/

Summary

Pick one verification mechanism, such as network taps, trusted execution environments, chip telemetry or satellite monitoring, and write a short, clear analysis of how it works and how it could fail.

Why it matters

International agreements on AI are only as good as our ability to check them. Very few people understand these mechanisms; good explainers help policymakers and new researchers.

What you will have at the end

A 2,000-word shallow dive in the style of Lennart Heim's blog.

How to start (tick each box as you go)

- [ ] **1. Read the summary of "Six Layers of Verification" and pick one mechanism.**
- [ ] **2. Read two or three of the specific papers on it from the verification reading list.**
- [ ] **3. Explain: what it checks, what it needs to be trusted, how a cheater might evade it, and how mature it is.**
- [ ] **4. Add a simple diagram.**
- [ ] **5. Ask someone technical to check it, then publish.**

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).

Compute
The raw computing power (specialised chips running for a long time) used to train and run AI models. Easier to measure and track than data or ideas.

Source and links

Technical AI Governance resource pack (MIRI TGT reading lists)
https://www.lesswrong.com/posts/GRPswd5v7haob97fA/recommendations-for-people-getting-into-technical-ai

Verifying International Agreements on AI: Six Layers of Verification
https://arxiv.org/abs/2507.15916

MIRI — Mechanisms to Verify International Agreements About AI Development

https://techgov.intelligence.org/research/mechanisms-to-verify-international-agreements-about-ai-development

Lennart Heim’s blog (examples of shallow dives)
https://blog.heim.xyz/

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