Stress-test one article of a proposed international AI agreement

Intermediate · A few weeks · Free · Project idea

Online version: https://ai-safety-sandbox.vercel.app/projects/stress-test-an-agreement-article/

Summary

MIRI TGT published a draft international agreement to prevent premature superintelligence. Pick one article and ask: how would it be verified, enforced and evaded?

Why it matters

Detailed proposals improve fastest when people outside the team attack specific clauses with care.

What you will have at the end

A memo critiquing one article and proposing improvements.

How to start (tick each box as you go)

- [ ] **1. Pick a decision-maker and a decision they actually face.**
- [ ] **2. Gather the facts: what exists, what is proposed, who disagrees.**
- [ ] **3. Lay out 2–3 realistic options with their pros and cons.**
- [ ] **4. Recommend one, and say what evidence would change your mind.**
- [ ] **5. Keep it to two pages and get feedback from someone who disagrees.**

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).

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

An International Agreement to Prevent the Premature Creation of ASI
https://arxiv.org/abs/2511.10783

Blog posts explaining parts of the agreement
https://www.lesswrong.com/s/238e8bGMPMuyZQQrT

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