The Commitment Atlas

Mapping AI red lines and testing whether they can be verified.

SPAR Fall 2026 project proposal. Mentor: Aryan Agarwal. July 2026.

Summary

Mentees build the Commitment Atlas: a public, sourced dataset of every AI red line commitment made by a government, company, or international process. The Atlas shows where commitments overlap, which are precise enough to test, and which carry any consequence. Two case studies then convert one commitment each into a draft testable standard. The outputs feed the founding design of the Touchstone Council, an independent verification institution now being built.

The problem

Red lines are multiplying. IDAIS Beijing proposed four in March 2024: no autonomous replication, no power seeking, no weapons of mass destruction uplift, no autonomous cyberattacks. Twenty companies signed the Seoul Frontier AI Safety Commitments. The Global Call for AI Red Lines gathered over 300 signatories at the UN General Assembly in September 2025. It demands enforceable red lines by the end of 2026. The Athens Roundtable identified seven red line domains in December 2025.

Almost none of this can be checked as written. On bioweapons uplift, Anthropic sets its threshold at basic STEM knowledge, OpenAI at novice actors, Amazon at expert level. Those are order of magnitude disagreements about what counts as unacceptable. Companies keep about half of their voluntary safety pledges. The Future of Life Institute graded seven major developers in 2025; none scored above C+. Most safety assessments are still done by the companies being assessed (UN Scientific Panel on AI, July 2026).

What exists and what is missing

METR maps common elements of company safety frameworks. SaferAI rates published risk management practices. The Future of Life Institute grades developers. All three cover companies only.

No public dataset spans state, corporate, and international red line commitments. None codes commitments for whether anyone could verify them. That is the gap this project fills.

Research questions

1. What has actually been committed? By whom, in what instrument, with what precision?
2. Where is the overlap? Which red lines do all major labs, and both the US and China, already accept?
3. Can the strongest commitments be converted into standards someone could test? What evidence and access would testing need?

What we will build

The Atlas. Every entry codes one commitment against a fixed codebook. Draft codebook v0, to be revised in week two:

- Actor and actor type: state, company, multilateral body, scientific assembly.
- Instrument: law, regulation, declaration, corporate framework, pledge.
- Date and status: active, superseded, lapsed.

- Domain, mapped to the seven Athens Roundtable clusters.
- Trigger type: capability, behaviour, or use.
- Threshold precision: quantified, defined, or vague.
- Verification hook: who could check, and what access checking would need.
- Consequence: none, internal halt, external sanction.
- Source link and exact quotation.

Case studies. Two commitments from the overlap set, one per study. Each drafts a candidate testable standard: definitions, evidence requirements, and plausible checkers. Each names what cannot be resolved on paper.

Precedent review. A short memo on FATF mutual evaluations, WADA lab accreditation, INPO peer reviews, and maritime load lines. It states what transfers to AI and what does not. It treats FATF's paper compliance record as a warning, not a footnote.

Twelve week plan

- Weeks 1 and 2. Onboarding. Codebook v1. Pilot coding of 15 commitments by two coders. Agreement measured, codebook revised.
- Weeks 3 to 6. Full collection and coding. Weekly adjudication of hard cases.
- Weeks 7 and 8. Analysis: the overlap set, precision and consequence distributions. Findings note drafted.
- Weeks 9 and 10. Case studies drafted. Precedent review completed.
- Week 11. External review through my network. Revisions.
- Week 12. Atlas v1, findings note, and case studies published. Synthesis memo delivered to the Council design process.

Deliverables

- Atlas v1: public dataset plus codebook.
- Findings note, around 3,000 words.
- Two case studies, around 2,000 words each.
- Precedent review memo, around 2,500 words.

All outputs carry named mentee credit.

Team design

Three mentees, four at most. Eight hours per week each.

- Seat 1: state and international commitments.
- Seat 2: company commitments. Familiarity with model evaluations helps.
- Seat 3: precedents and standards, co-drafting the case studies.

A fourth mentee makes the case studies a separate seat.

How I work as a mentor

One weekly team call. A short weekly one to one per mentee. Written feedback on every draft, every week. Roughly two hours per mentee per week in total. Mentees make the coding calls in their workstream and defend them in adjudication. Nothing publishes without joint sign off.

About the mentor

I am a policy analyst at the OECD in Paris, working on AI and research policy. I co-led ministerial negotiations for the Global Partnership on AI. I previously co-founded a regulated fintech. I am building the Touchstone Council and mentor in a personal capacity.

Risks and honest limits

- Atlases can reward paper commitments. The Atlas therefore describes and codes. It grades nothing. Grading is a later institutional job.
- Coding drift. Mitigated by the pilot, measured agreement, and weekly adjudication.
- Scope creep. Frontier AI red lines only. General AI ethics principles are out.
- Access. Public documents only. No confidential material is used or sought.

Relationship to prior work

This project builds on Thwing et al. 2026, the Arcadia Impact study of the international AI measurement network. That study argues the intergovernmental Network should grow into the verification role. The Touchstone Council concept shares the diagnosis and departs on the prescription: an independent body, funded by no one it assesses. I have no authorship role in that study. Moreira Tomei, Jain, and Franklin 2025 map four market channels that could give findings consequence without a treaty: insurance, audit, procurement, and disclosure.

Selected sources

- Thwing, Gringras, Richard, and Tang (2026). Strengthening the International Network for Advanced AI Measurement, Evaluation and Science to Support Enforceable Global AI Red Lines. Arcadia Impact. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6854278
- Moreira Tomei, Jain, and Franklin (2025). AI Governance through Markets. https://arxiv.org/abs/2501.17755
- Global Call for AI Red Lines (2025). https://red-lines.ai
- IDAIS Beijing statement (2024). https://far.ai/post/2024-03-idais-beijing/
- METR (2025). Common elements of frontier AI safety policies. https://metr.org/blog/2025-12-09-common-elements-of-frontier-ai-safety-policies/
- Future of Life Institute (2025). AI Safety Index, Summer 2025. https://futureoflife.org/ai-safety-index-summer-2025/
- The Future Society (2026). Seventh Athens Roundtable outcomes.
- UN Scientific Panel on AI (July 2026). Finding on developer conducted safety assessments.