Trust, But Verify

What Five Historical Treaties Teach AI Diplomats

Tara L. Thwing, Alixandra Acker, Joel Christoph, Uma Kalkar, Julian Theseira · Arcadia Impact AI Governance Taskforce · 19 August 2026

Summary

The problem: No state or bloc with real influence over advanced AI has accepted a binding international compliance system.

Why it matters: AI governance is proliferating into summits, coalitions, and national laws pulling in different directions: fragmentation that can look like progress while holding no one accountable. A large-scale meta-analysis of international treaties found most fail to achieve their intended effects, and identified enforcement mechanisms as the one design choice with real potential to change that.

The argument: Verification only works when the governed activity can be measured from outside, and it works best when the party being verified helped design the mechanism rather than being handed a finished one.

Recommendations: Agree on a measurable proxy before arguing over treaty text, bring AI developers into designing the verification regime, and structure obligations so no party bears a disproportionate burden.

A Preview of How Fast This Moves

In June 2026, the U.S. government suspended access to Anthropic's most capable models, Claude Mythos 5 and Fable 5, on national security grounds. Three weeks later, it reversed course. Whatever one makes of the decision itself, the episode was a preview: frontier AI risk moves across borders and through global infrastructure faster than any government can respond to it, and right now the response is mostly improvised, one country at a time.

That is the governance gap. There is broad agreement that advanced AI needs international oversight, and far less agreement on what it should require. The U.S. has built capability-based coalitions like Pax Silica while imposing unilateral export controls. China has pushed a

UN-centered approach while also standing up the World Artificial Intelligence Cooperation Organization outside it. The EU has the only binding, cross-sector AI law in force, and is now softening enforcement under competitiveness pressure. Summits and dialogues have produced shared vocabulary but no binding text: a lot of activity, very little that would hold a state accountable for breaking a rule.

Three Findings Stand Out

Diplomats will eventually design a real compliance mechanism, and when they do, most of the available evidence about what works comes from other domains: arms control and environmental treaties. We compared five of them: the Nuclear Non-Proliferation Treaty (1968), the Biological Weapons Convention (1972), the Chemical Weapons Convention (1993), the Montreal Protocol (1987), and the Paris Agreement (2015). Each was scored against the same three-part test: does the rule bind and can conduct be judged against it, can someone other than the assessed state verify compliance, and do any defined consequences follow a breach. The third question carries significant weight: the same meta-analysis that found most treaties fail to achieve their intended effects identified enforcement as the one design choice with real potential to change that.

First, political conditions set the ceiling on what a treaty could achieve, before anyone drafted a clause. The Non-Proliferation Treaty secured broad participation by offering civilian nuclear access in exchange for a freeze on recognized weapons states, compensation for an obligation that would otherwise have fallen unevenly on non-nuclear states. Montreal's Multilateral Fund played the same role for developing countries facing binding phase-out schedules. Paris took the opposite route to the same goal: instead of compensating reluctant states for a strong obligation, it made the obligation itself optional and nationally set. Different treaties, same constraint: negotiators only ever chose from what the room allowed.

Second, verification only worked where the governed activity could be measured from outside, independent of what the assessed state chose to report.

> *"Agreements built around a state's declared purpose or intent, like the Biological Weapons Convention's ban on agents lacking a peaceful justification, could never be verified: no inspectorate was ever built to check it. Agreements built around a concentrated, countable input, like the handful of firms producing ozone-depleting chemicals under the Montreal Protocol, could be checked from trade data alone, without a single on-site inspection."*

Third, whether private industry helped design the verification mechanism, rather than being presented with a finished one, shaped whether verification actually worked. Chemical manufacturers helped draft the Confidentiality Annex that let inspectors verify compliance while protecting trade secrets, and the Chemical Weapons Convention entered into force with a working inspectorate. Pharmaceutical and biotech firms had no comparable role in the Biological Weapons Convention's design; when a verification protocol was finally negotiated decades later, industry resisted it, and the U.S. rejected it in 2001. The same commercial concern about proprietary information produced opposite outcomes depending on when industry was brought in.

Where This Lands for AI

Among the inputs to advanced AI (compute, algorithms, and data), compute currently behaves like Montreal's chemicals: physical, chokepointed, and traceable through a small number of suppliers. That is a real, usable advantage, and not a permanent one. As algorithmic efficiency improves, a fixed training-compute threshold becomes a worse proxy for capability, since the same capability is increasingly reachable with less compute. Verification aimed at inference (tracking resource use at deployment, not just at training) is less exposed to that erosion, though neither approach alone is a durable solution.

That leaves negotiators with a narrow, closing window and a fairly specific job.

Recommendations

Agree on a measurable proxy, and invest in the capacity to verify it. Compute is the most plausible starting point: physical, chokepointed, and traceable, the same properties that made Montreal's chemicals checkable from trade data alone. Invest in that capacity now, while compute is still concentrated, rather than improvising a method inside a short negotiating window.

Bring developers into designing the verification regime, not just the outcome. As the CWC-BWC contrast above shows, early co-design is what separated a working inspectorate from a stalled one. It's no guarantee by itself, though: it worked partly because inspection was commercially convenient for the firms involved, and frontier AI developers face the opposite incentive, since revealing capability to rivals or regulators carries a cost with no offsetting upside.

Structure obligations so no party bears a disproportionate burden. Nearly every developing country ratified Montreal only once the Multilateral Fund offset the cost of a phase-out schedule that would otherwise have limited their industrial development unfairly relative to major emitters, and that compensation was written in from the outset, not added under pressure. Further study is needed to document the elements of effective benefit-sharing mechanisms.

The Window Will Not Stay Open

None of this waits on a comprehensive treaty becoming politically possible. The advantage AI governance has right now, a measurable and concentrated input, is the same advantage that made Montreal work, and it is already eroding. Settle on what can be measured, build the capacity to measure it, and design the verification regime with the parties it will apply to, before the politics that make a real agreement possible arrive and find the technical groundwork still undone.

About the authors

Tara L. Thwing, Alixandra Acker, Joel Christoph, Uma Kalkar, and Julian Theseira are researchers with Arcadia Impact's AI Governance Taskforce, a policy research programme for professionals transitioning into AI governance careers. This post draws on their policy brief, "Designing Verification and Enforcement for International AI Governance: Lessons from Historical Treaties," produced with advisory input from José J. Villalobos at the Future of Life Institute.

This post represents the views of its authors rather than the views of Arcadia Impact or any partner organization.