

The Financial Action Task Force Has Some Lessons for Frontier AI Governance - Are Governments Ready to Hear Them?

Tara Thwing, David Gringras, Ulysse Richard, Matthew Tang

Arcadia Impact AI Governance Taskforce

May 2026

In 1989, G7 finance ministers gathered in Paris and, facing a growing drug-money crisis, created a new body to address it. They gave it sixteen members, a one-year mandate, no treaty, no enforcement authority, and no democratic mandate beyond the political will of the room. Today, the [Financial Action Task Force's \(FATF\)](#) standards are implemented by over 200 jurisdictions. By any conventional theory of how international institutions work, this shouldn't have happened.

The lessons we draw from this puzzle are directly relevant to one of the hardest problems in AI governance right now.

The conventional wisdom holds that meaningful international coordination requires either a formal treaty (slow, and perhaps impossible given US-China dynamics) or coercive enforcement (politically toxic in any configuration that matters). FATF's existence is a near-four-decade empirical refutation of that assumption. The question worth asking is not whether we can replicate FATF's outcome, but what specific design choices made that outcome possible, and whether they transfer to AI.

Four major things explain FATF's standard-adoption success, and none of them is direct legal authority or a treaty.

Lesson 1: Principles-based standards designed to travel.

FATF's original Forty Recommendations specified what member states should achieve, not how exactly their domestic law should achieve it. FATF deliberately acknowledged that countries have diverse legal and financial systems and so all cannot take identical measures. That flexibility is a large part of what allowed standards developed by sixteen G7 nations to become operative across 200 jurisdictions without requiring any of them to wholesale rewrite their legal architecture. Once a common framework exists, deviations become legible as non-compliance rather than reasonable disagreement. Frontier AI governance does not yet have this. Twelve major AI companies have published safety frameworks, but their risk threshold definitions diverge substantially even for identical threat categories, and those are not cosmetic differences in phrasing. They define fundamentally different objects of concern.

Lesson 2: Regional architecture that extends both reach and legitimacy.

FATF's nine regional bodies brought over 160 additional jurisdictions into the evaluation framework, as participants in implementation and peer review rather than passive recipients of standards written elsewhere. This is what made near-universal adoption legitimate rather than merely widespread. The [International Network for Advanced AI Measurement, Evaluation and Science](#), the international body coordinating AI safety institutes, currently has its membership heavily concentrated in Western and high-income Asian states, with Kenya the sole African affiliate. A compliance signalling mechanism deployed before regional architecture confers legitimacy will reproduce exactly the selective enforcement critique that collapsed FATF's original blacklist.

Lesson 3: sequencing discipline.

This is where the measurement science matters. FATF spent its first decade building standards and evaluation credibility before introducing any enforcement mechanism. When it broke that discipline, launching the Non-Cooperative Countries and Territories blacklist in 2000 before the procedural legitimacy to sustain it was in place, the mechanism collapsed within six years. Read structurally, the failure was sequential: findings cannot be consequential before they are credible, and they cannot be credible before the standards they are measured against are legitimate. For the Network, the sequencing constraint is partly procedural and partly scientific. The evaluation science itself is not yet reliable enough to sustain threshold-based consequences; [format choice alone shifts measured safety scores by five to twenty percentage points across identical items](#). That is the argument for treating this phase as foundation-laying rather than enforcement. The pressure on the Network to skip that sequence should be resisted with some urgency.

Lesson 4: market amplification as a substitute for legal compulsion.

FATF has never had the direct authority to compel anyone to do anything. What it has is a grey and black list mechanism whose publication drives de-risking behaviour by correspondent banks, imposing costs through decentralised market decisions that FATF itself never had the authority to levy directly. The compliance effect depends on sustained belief in the mechanism's credibility, which is itself a product of the sequencing and legitimacy work that preceded it.

Each of these has a direct analogue in the Network's current or potential architecture. The Network is not treaty-based, which is a feature rather than a bug, because it allows the Network to operate at the pace that a domain evolving faster than conventional international law requires. Its emerging evaluation methodology could form the basis for common standards. Publication of evaluation findings could generate reputational consequences that function as market incentives. The OECD's existing [AI Policy Observatory](#) offers an administrative hosting arrangement that mirrors the institutional move FATF made in 1989.

Two disanalogies deserve honest acknowledgement. FATF evaluates sovereign states, whose legal architecture does not reconfigure itself when an assessment team arrives. AI models can.

And FATF's market amplification mechanism works because the global financial system is interconnected enough that listing imposes costs regardless of a jurisdiction's preferences. The AI market is bifurcating along US-China lines in a way global finance never did, which means listing would work within geopolitical blocs rather than across them.

Both disanalogies sharpen the design constraints rather than removing the model from consideration.

The full paper works through what that design would concretely require. The starting point is this: the architecture that made FATF consequential is available to the Network with existing resources and no new legal authority. A formal treaty may not be forthcoming, so the remaining question is whether the foundational institutional work that doesn't need one gets done. The Network will also require sustained investment, political will, and trust; whether member governments are willing to make that commitment in this arena is an open question.

Read the full paper [here](#).

Acknowledgment: The authors gratefully acknowledge Su Cizem, Niki Iliadis, and colleagues at The Future Society for generously sharing their expertise and providing substantive feedback during the development of this work. The analytical conclusions and policy recommendations of this study are those of the Arcadia Impact AI Governance Taskforce authors alone. Special thanks are also due to Peter Courtney, the initial Team Lead of this study.