
AI Treaty Momentum Index

Oksana Kotelnikova
The AGI Collective

Sofia Lobanova
NLP Data Scientist

With
Apart Research

Abstract

We present the AI Treaty Momentum Index (ATMI), a lightweight forecasting model that estimates the near-term **chance of a new international AI risk treaty or binding regulation**. ATMI learns from past treaties in climate, nuclear, biosafety, and finance by encoding each case with **mechanism classes—buckets of factors** that help or hinder agreement (e.g., Policy window, Capability shock, Verification feasibility, Public salience, Epistemic shift, Geopolitical alignment). Each mechanism is scored from **−5 (hurts)** to **+5 (helps)** per case, and we map these mechanism profiles to outcomes (treaty vs. no treaty).

We use the fit in two modes: **Nowcasting**—apply today’s AI governance signals to produce a current chance with 80% intervals; **Scenario-casting**—stress-testing futures where scenarios fork from AI-2027 milestones to see how the odds shift.

Early findings: the largest marginal lifts come from Epistemic shift and Geopolitical alignment, followed by Public salience and Epistemic infrastructure.

As of Nov 2025, ATMI estimates a **~0.6% chance** of a new treaty, suppressed by misaligned geopolitics, weak verification readiness, and negative epistemic/resource shifts.

We propose to follow this MVP and build a public “**anti-doomsday**” tool: a web dashboard that tracks each mechanism with live indicators, shows the ATMI curve over time, and flags **actionable weakest links** (the marginal mechanisms whose improvement most raises the chance of agreement).

Keywords: forecasting, AI governance, early warning, scenario analysis, Bayesian logistic

1. Introduction

Question. Do stable patterns in triggers/enablers combinations and scopes predict treaty or regulation creation—and can we use them to forecast AI risk international treaty creation?

Why it matters. If we can quantify “momentum” toward a treaty, we can focus advocacy and diplomatic effort on the **bottleneck mechanisms**, shorten time-to-agreement, and reduce tail risks from rapidly scaling AI.

Threat model. Transformative/existential risks from fast capability growth: failure modes include multi-actor races with uncontained model capability development and insecure deployment.

Contribution. A compact, transparent index + dashboard that codifies historical regularities, nowcasts current momentum for AI safety coordination, and stress-tests future scenarios (branches built on AI-2027 milestones).

2. Methods

Data & schema

- **Case** = one attempt to create an international treaty or binding regulatory regime (success = entered into force; failure = credible window closed without adoption).
- **Condition** = one **trigger** (time-bounded change) or **enabler** (slow/institutional prerequisite). Coded with direction (+/-) and intensity scope (0-5).
- **Mechanism classes:** *Policy window, Verifiable threshold ready, Epistemic infrastructure, Information sharing, Verification feasibility, Feasibility path for industry, Finance vehicle ready, Geopolitical alignment, Legal plumbing, Capability shock, Incident cluster, Epistemic shift, Public salience, Resource shift.*
- **Aggregation** per case $\times$ mechanism class. Negative direction items (e.g., *Geopolitical misalignment*) enter as negative scope.

Model

- We fit a Bayesian logistic regression to historical treaty cases with standardized mechanism-class scopes and weakly-informative priors ($\beta_j \sim N(0,1)$), reporting posterior means with 80% credible intervals. We aimed to estimate priors for mechanism classes and use them to compute the probability of an AI alliance/binding regulations for the present and future scenarios. Our current logistic-regression model does not yet incorporate these priors; we will align the model logic in future iterations.
- **Cause-area slices:** climate/nuclear/bio/finance (historical) vs. AI (current/future scenarios).

Logistic regression with group-specific intercepts for probability of AI international treaty/regulation.

We fit a logistic regression model with shared mechanism coefficients and random intercepts per cause area, estimated via maximum likelihood.

The model predicts the log-odds of success using:

$$\text{logit}(p) = \text{intercept_for_cause_area} + \sum_{\text{mechanism_value_m}} (\text{mechanism_value_m} \times \text{beta_m})$$

- No priors or sampling: parameters are estimated via L-BFGS optimization (MLE).
 - Intercepts (α_j) capture baseline success likelihood for each cause area.
 - Mechanism weights (beta_m) are shared across all areas and represent global influence.
 - Feature normalization and L2 regularization are applied to prevent overfitting.
- **Out-of-sample prediction for novel cause areas.**

For previously unseen cause areas, we approximate their intercept using the **mean intercept** from observed areas. This allows prediction transfer:

$$\text{predicted_p} = \text{sigmoid}(\text{mean_intercept} + \sum_{\text{mechanism_value_m}} (\text{mechanism_value_m} \times \text{beta_m}))$$

- **Robustness.**
 - **Norm-based constraint:** L2 regularization is applied to both intercepts and mechanism weights, constraining extreme values when data is sparse.
 - **Input standardization:** All mechanism features are normalized to have zero mean and unit variance.
 - **Transfer test ready:** The model can be evaluated via slicing by cause area (e.g. climate, bio, AI) or leave-one-area-out tests.

Trigger mechanism taxonomy (observable, non-gameable where possible)

- **Capability shock** = Unexpected AI advances or setbacks that rapidly shift urgency (trigger)
- **Epistemic infrastructure** = Standing institutions and standards producing shared, trusted evidence (enabler)
- **Epistemic shift** = Shifts in expert consensus or framing that redefine problem (trigger)
- **Feasibility path for industry** = Credible, affordable compliance pathways and substitutes for affected firms (enabler)
- **Finance vehicle ready** = Dedicated funding mechanisms exist to finance implementation and transitions (enabler)
- **Geopolitical alignment** = Pivotal states sufficiently aligned to cooperate and share costs (enabler)
- **Incident cluster** = A cluster of salient incidents concentrates leaders'/media attention and raises public awareness (trigger)
- **Information sharing** = Channels and norms for timely disclosure and learning (enabler)
- **Policy window** = Time-bound political opportunity and mandate to advance agreement (enabler)
- **Public salience** = Sustained public/media attention creating pressure for action (trigger)
- **Resource shift** = Changes in resources or leverage that reshape bargaining power (trigger)
- **Venue + depositary** = Suitable forum and legal home identified to host treaty (enabler)
- **Verifiable threshold ready** = Clear, accepted thresholds and targets defined for obligations (enabler)
- **Verification feasibility** = Practical, credible methods exist to monitor and enforce compliance (enabler)

Nowcast & scenarios

- **Nowcast (Nov 2025):** code present-day indicators for each mechanism (conservative scopes). Map to mechanism classes.
- **Scenario-cast** with AI-2027 forks:
 - **Milestone 1** (*early-2026*): *Policy rumblings*

Context add-on to AI-2027: Immediately after Agent-1's public release (early 2026), major parties in key jurisdictions roll out AI-regulation planks. Track-1.5 dialogues appear (US–EU; US–MENA). Scientists get some stage time but do not shift political direction. The US and EU stand up separate committees to model AI x-risks and mitigations, with light expert exchange.

- **Milestone 2** (*late-2026*): *Job disruption & misinfo surge*

Right after Agent-1-mini (10× cheaper) floods the market and “AI takes some jobs”. New dynamics: youth unemployment rises; major unions begin coordinating with PauseAI-type movements; developed countries see upticks in crime/drug use. Large-scale mis/disinfo and deep-fake campaigns accelerate.

- **Milestone 3** (*Apr-2027*): *Agent-3 leaks + Taiwan war & chip shock*

As OpenBrain works to align Agent-3, leaks reach the public and the PauseAI movement becomes electorally salient. War erupts in Taiwan, triggering global chip shortages. New US tariffs ripple through metals/solar/power, and leading US AI companies slow marginally.

3. Results

From a corpus of historical treaty/regulation episodes (34 total, 24 successes/10 failures), we coded conditions into standardized **mechanism classes** and estimated class-level **priors** (log-odds effects on success) using a Bayesian logistic model. Coding and synthesis were supported by GPT-5 Thinking.

Figure 1 – Influence (priors) table (log-odds impact, standardized)

mechanism_class	prior median	prior range	notes
Policy window	0.10	[-0.15, 0.34]	n_active $\geq$ 2: 3 (SR=0.67); area: Climate & env. (5.0, n=2)
Verifiable threshold ready	-0.20	[-0.45, 0.00]	n_active $\geq$ 2: 2 (SR=0.50); area: Climate & env. (3.0, n=2)
Capability shock	0,3	[0,00; 0,48]	n_active $\geq$ 2: 0; scope $\rightarrow$ SR: neg:0.00 (no strong direct effect in this sample)
Epistemic infrastructure	0,57	[0,25; 0,87]	n_active $\geq$ 2: 15 (SR=0.93); area: (Arms control 3.0, n=4); scope $\rightarrow$ SR: 4+:0.93, 2-3:1.00, neg:0.00
Epistemic shift	0,84	[0,58; 1,09]	n_active $\geq$ 2: 14 (SR=0.93); area: Arms control (7.25, n=8); scope $\rightarrow$ SR: 4+:1.00, 2-3:0.83, neg:0.00
Feasibility path for industry	-0,24	[-0,42; 0,00]	n_active $\geq$ 2: 4 (SR=0.75); area: Climate & env. (1.0, n=3)
Finance vehicle ready	0,1	[-0,05; 0,26]	n_active $\geq$ 2: 4 (SR=1.00); scope $\rightarrow$ SR: 4+:1.00, 2-3:1.00 (small N)
Geopolitical alignment	0,77	[0,51; 1,01]	n_active $\geq$ 2: 6 (SR=0.83); area: Climate & env. (8.0, n=2); neg scope cases (n=7) had SR=0.00
Incident cluster	0,49	[0,27; 0,73]	n_active $\geq$ 2: 10 (SR=1.00); area: Arms control (7.75, n=5); scope $\rightarrow$ SR: 4+:1.00, 2-3:1.00, neg:0.00
Information sharing	0,02	[-0,31; 0,27]	n_active $\geq$ 2: 4 (SR=0.75); area: Climate & env. (5.0, n=2)
Policy window	0,17	[-0,10; 0,48]	n_active $\geq$ 2: 3 (SR=0.67); area: Climate & env. (5.0, n=2)
Public salience	0,62	[0,31; 0,93]	n_active $\geq$ 2: 16 (SR=0.88); area: Maritime law (6.0, n=3); scope $\rightarrow$ SR:

			4+:—, 2-3:—, neg:— (bins sparse)
Resource shift	0,53	[0,24; 0,71]	n_active $\geq$ 2: 1 (SR=1.00); scope $\rightarrow$ SR: 4+:1.00, neg:— (very small N; treat as noisy)
Venue + depositary	0,16	[-0,22; 0,52]	n_active $\geq$ 2: 16 (SR=0.88); area: Arms control (5.8, n=7)
Verifiable threshold ready	-0,38	[-0,55; 0,00]	n_active $\geq$ 2: 2 (SR=0.50); area: Climate & env. (3.0, n=2)
Verification feasibility	0,5	[0,17; 0,84]	n_active $\geq$ 2: 7 (SR=1.00); area: Arms control (6.4, n=7)

Highlights

Hard preconditions dominate: Epistemic shift and Geopolitical alignment are the strongest single predictors, followed by Public salience and Epistemic infrastructure. Incident cluster, Resource shift, and Verification feasibility are close behind.

Venue + depositary is a force multiplier, not a prime mover: Solo effect is modest ($\times 1.17$ per +1), but it tends to be present in winning packages because it lets other mechanisms “cash out” into a tractable instrument.

Asymmetry matters: Negative scope on Geopolitical alignment, Epistemic shift, or Verification feasibility is highly predictive of failure in the scenarios we scored (large negative log-odds contributions when scope < 0).

Design implication: Prioritize flipping the sign (from negative $\rightarrow$ neutral/positive) on the top priors before trying to push already-positive but lower-leverage enablers.

Combos that “nearly guarantee” success

The **highest-leverage packages** when each mechanism is at **aggregate scope ≥ 2** are:

- **Epistemic shift + Geopolitical alignment**, ideally with **Public salience and Venue + depositary** in place.
- **Incident cluster + Public salience**, converted into policy via **Venue + depositary** and backed by **Verification feasibility**.
- **Epistemic infrastructure + (Epistemic shift or Incident cluster)**, plus **Venue + depositary**.

These packages systematically outperform a neutral baseline in our case scoring. Exact success rates by combo should be re-tabulated on the full historical matrix; the priors indicate direction and relative lift.

Cause-area differences

- **Arms control:** Historically aligns with strongly positive priors for **Verification feasibility**, **Incident cluster**, **Epistemic shift**, and benefits from **Venue + depositary**.
- **Climate & environmental governance:** Greater reliance on **Geopolitical alignment**, **Policy window**, and **Incident cluster**; **Public salience** can be volatile.
- **Maritime/other regimes:** Often hinge on **Epistemic shift** plus **Public salience**, then convert via **Venue + depositary**.

These are directional signals; small Ns in some areas mean we should treat them as hypotheses to test rather than firm estimates.

Limitations and next steps

Findings are interpretable but preliminary: the dataset is small, class definitions are still coarse, and the mapping of triggers/events to mechanisms needs tightening.

We should re-run on a larger corpus with sharper taxonomy and adjudicated coding, and explicitly model **blocker chains** (e.g., weak information sharing → fewer incident disclosures → slower public awareness/epistemic shift) to estimate their correlation/mediation effects.

Nowcast

Probability of adopting an **international AI treaty or binding regulation**—**0.6%**

Drivers: Feasibility path for industry, Verifiable threshold ready.

Brakes: Resource shift, Epistemic shift, Verification feasibility, Geopolitical alignment.

Advice (how to raise odds for international treaty/regulations):

- **Verification feasibility & infrastructure:** stand up a minimally-viable *joint* evals bench; flip scopes toward neutral. Currently blocked by **Information sharing** (negative).
- **Resource shift:** earmark visible budget for incident reporting & evals; even small +1–2 scope change helps.
- **Geopolitical alignment:** create a minilateral lane (US-EU-UK-JP) with a dated workplan and depositary path; aim to reopen a policy window (currently slightly negative).

Forecasting AI scenarios

AI_F1 (early 2026): Policy rumblings

Probability of adopting an **international AI treaty or binding regulation**— **0.83%**.

Context add-on to AI-2027: Immediately after Agent-1’s public release, major parties in key jurisdictions roll out AI-regulation planks. Track-1.5 dialogues appear (US–EU; US–MENA). Scientists get some stage time but do not shift political direction. The US and EU stand up separate committees to model AI x-risks and mitigations, with light expert exchange.

Drivers: Incident cluster, Feasibility path, Policy window.

Brakes (negatives): Epistemic shift, Geopolitical alignment, Epistemic infrastructure.

Advice (how to raise odds for international treaty/regulations): orchestrate a multi-stakeholder “risk consensus” & depositary-ready draft to move epistemics and alignment toward neutral.

AI_F2 (late 2026): Job disruption & misinfo surge

Probability of adopting an **international AI treaty or binding regulation— 0.63%.**

Right after Agent-1-mini (10× cheaper) floods the market and “AI takes some jobs” in the AI-2027 scenario. New dynamics: youth unemployment rises; major unions begin coordinating with PauseAI-type movements; developed countries see upticks in crime/drug use. Large-scale mis/disinfo and deep-fake campaigns accelerate.

Drivers: Incident cluster, Resource shift, Verification feasibility.

Brakes (negatives): Epistemic shift, Geopolitical alignment, Epistemic infrastructure.

Advice (how to raise odds for international treaty/regulations): publish a *joint* thresholds report (labs+academia+regulators) and pre-negotiate a lowest-common-denominator control set to reopen the policy window.

AI_F3 (Apr 2027): Agent-3 leaks + Taiwan war & chip shock

Probability of adopting an **international AI treaty or binding regulation—0.1%.**

As OpenBrain works to align Agent-3 (Apr 2027), leaks reach the public and the PauseAI movement becomes electorally salient. War erupts in Taiwan, triggering global chip shortages. New US tariffs ripple through metals/solar/power, and leading US AI companies slow marginally.

Drivers: Resource shift, Verification feasibility, Policy window.

Brakes (negatives): Geopolitical alignment (very large), Epistemic shift, Incident cluster.

Advice (how to raise odds for international treaty/regulations): pursue a coalition-of-the-willing treaty path with depositary pre-commitment to bypass full consensus; invest in structured incident reporting to flip *Incident cluster* and *Public salience*.

Outlier contributions across scenarios

- single mechanisms dominate their cases' odds.

- AI_F1: Epistemic shift - negative

Several **high-magnitude negative conditions** are present under this mechanism—e.g., “*High-profile failures reduce perceived urgency*” (−5), “*Competitiveness-only framing crowds out risk governance*” (−5), plus “*All-clear narratives*” (−2) and “*Attribution disputes/politicization*” (−3). The few positives (e.g., expert letters, reinsurer signals) are smaller, so the **net signed scope is strongly negative** and dominates the case.

- AI_F2: Epistemic shift - positive

New positive enablers outweigh earlier negatives. **Negatives:** “all-clear” narratives −2, politicization/over-claiming −3, competitiveness-only framing −5; **Positives:** reinsurer exclusions +3, eval/red-teaming commitments +3, expert letters on an “international rules gap” +4, capex signals +1, debates on misgeneralization/deception +2.

- AI_F3: Geopolitical alignment - negative

Three **max-severity alignment shocks**—*Taiwan war disrupts alignment* (−5), *AI-as-strategic-race narrative hardens* (−5), and *escalating trade controls deepen bloc rivalry* (−5)—with only a small positive (*allies propose a compute stewardship regime*, +1).

Limitations

- **Sparse & collinear bins:** Some mechanisms (e.g., **Finance vehicle ready**, **Policy window**, **Verifiable threshold ready**) have few active instances and are correlated with higher-leverage classes. Interpret intervals cautiously.
- **Associational, not causal:** Priors come from observed associations; strong-looking packages may proxy for unobserved context (e.g., domestic politics).
- **Small scenario panel for correlations:** The 4 recent cases show very high co-movement (e.g., **Capability shock** ↔ **Information sharing** nearly perfectly anticorrelated), which can distort combo inferences. Expand cases to stabilize estimates.

4. Discussion and Conclusion

What actually moves treaties/regulations:

The biggest drivers in our data are **epistemic shift** (the expert consensus hardens), **epistemic infrastructure** (credible institutions and standards in place), **public salience** (sustained attention), and **incident clusters** (several related events close together). **Geopolitics: negative alignment kills** (all negatives failed in our sample).

Caveats:

Some bins are small, effects are associations (not proof of causality), and our current fit doesn't model timing yet.

Recommended actions - for policy teams, funders, civic leaders, standards bodies, journalists:

1. **Consolidate epistemics:** credible evaluators + visible endorsements.
2. **Plumbing on the shelf:** ready-to-file clauses + adoption pathway.
3. **Salience with substance:** coverage tied to concrete proposals.
4. **Manage geopolitics:** build track-1.5/2 bridges to avoid hard splits.

Approaches to test with the follow-up research:

- Qualitative Comparative Analysis (QCA). We should search for proven condition combinations and blocker groups: literature review reveals multiple studies pointing to **trigger-enabler combinations/sequences** that raise the likelihood of international treaty/regulation setup. [1-5]
 - Process tracing of pivotal cases: to test whether hypothesized mechanisms actually fired (not just post-hoc stories).
 - Policy window modeling (Kingdon): matching problem stream (incidents & evidence), policy stream (ready text), politics stream (leadership windows).
 - Early-warning indicators & change-point detection: flag rising "momentum" (e.g., evaluator adoption + insurer behavior + draft clauses surfacing).
 - Structured expert elicitation (Delphi/crowd forecasting): priors on which bundles matter and expected lead times.
1. **Time-weighting:** Add **prior decay by recency** for triggers (e.g., weight = 1 at same month; monotone decay with lag).
 2. **Combo tables on the full matrix:** Re-compute success rates by scope thresholds (≥ 2 , ≥ 3 , ≥ 4) for top combos using the updated priors.
 3. **Ablations:** Re-fit with each top mechanism dropped to quantify dependence on **Epistemic shift/Geopolitical alignment**.
 4. **MC for intervals:** Simple Monte Carlo by sampling priors within their ranges and scope within observed bins $\rightarrow$ 80% intervals on case probabilities and combo SRs.

5. References

1. Keohane, Robert O., and David G. Victor. 2011. "The Regime Complex for Climate Change." *Perspectives on Politics* 9(1): 7–23. Available at: https://ilar.ucsd.edu/files/publications/studies/2011_the-regime-complex.pdf
2. Benedick, Richard E. *Ozone Diplomacy: New Directions in Safeguarding the Ozone Layer*. Cambridge, MA: Harvard University Press. Available at: <https://www.jstor.org/stable/j.ctv1smjv7m>
3. Fuhrmann, Matthew, and Yonatan Lupu. 2016. "Do Arms Control Treaties Work? Evidence from the NPT." (Resource packet on NPT effectiveness & safeguards upgrades, incl. IAEA Model Additional Protocol INFCIRC/540 and Hirsch, 2004.) Available at: https://www.odvv.org/resources/attachment/1653199467_73f681afb7c03e0af45dec0b328d5311.pdf
4. Congressional Research Service. 2014. *The Evolution of Cooperative Threat Reduction (Nunn–Lugar): Issues for Congress* (CRS Report R43143). Available at: https://www.everycrsreport.com/files/20140613_R43143_038d1e15a739cce3b67f79608de94dd0b93798e0.pdf
5. Helleiner, Eric; G20/FSB documents; and Pagliari, Stefano. *FSB & Basel III—Creation and Design*. Waterloo, ON: CIGI. Available at: https://www.cigionline.org/static/documents/g20_no_1_2.pdf
6. Nordhaus, William D. 2015. "Are We Approaching an Economic Singularity? Information Technology and the Future of Economic Growth." NBER Working Paper No. 21547. Available at: https://www.nber.org/system/files/working_papers/w21547/w21547.pdf

6. Appendix

Security Considerations

- **Misuse risk:** Model could be gamed to stage/overhype triggers; mitigated by sourcing only high-credibility indicators and cross-verification.
- **Overconfidence:** Priors are subjective; mitigate with wide prior ranges, sensitivity analysis, and explicit caveats.
- **Data drift:** Policy processes change; schedule periodic recalibration.
- **Selection bias:** Western-centric sources; plan to add multilingual monitoring & regional experts.
- **Ethical guardrails:** No naming/shaming specific negotiators; focus on public, structural indicators.

Project proposal

Goal: Ship a public, minimal **AI Treaty Momentum Index** that shows today's odds, highlights the weakest link to fix, and lets you toggle future scenarios.

We turn governance noise into a simple **momentum meter** with clear levers: get plumbing + venue ready, lock in the epistemics, keep salience real - and watch the probability curve rise. Policy teams, funders, standards bodies, journalists.

1) Product

- **ATMI nowcast** with 80% uncertainty band (12–24 month horizon).
- **Mechanism tracker** for: Legal plumbing, Epistemic shift/infrastructure, Public salience, Incident cluster, Venue + depositary, Geopolitics, Verification feasibility.
- **Bundle meter:** flags when “win bundles” are close (e.g., Venue + Salience).
- **Weakest-link tip:** “+1 in X → +Y% success.”
- **Scenario toggles:** Agent-1 rumblings, job disruption, Agent-3 leaks + Taiwan conflict.

2) Data - what we'll add:

- Recent cases and non-Western data.
- Time/sequence (decay + “enabler before trigger?”).
- Negative signals/blocker groups.
- Indicators per mechanism (e.g., draft text exists; agenda item lodged; eval protocol accepted).
- Light double-coding for consistency.

3) Validation

- Backtest on held-out cases; **pre-register** monthly nowcasts; publish hits/misses.
- Red-team with policy + stats reviewers; quick user tests with policy teams.

4) Timeline

- **Weeks 1–2:** finalize schema, expand data, baseline model.
- **Week 3:** dashboard v1 (nowcast, bundles, weakest link).
- **Week 4:** scenarios + public write-up/demo.