
VERIFICATION MECHANISM FEASIBILITY SCORER¹

Moneera Yassien
Independent

Yatharth Maheshwari
Independent

Valeriia Povergo
Independent

Jayani Srinivasan
Independent

Erik J. Leklem
Independent

Alexandra Moraru
Independent

With
Apart Research

Abstract

Despite growing international interest in verifiable AI governance—from the 2023 Bletchley Declaration to the 2025 Paris AI Action Summit—treaty negotiators and policy makers lack a systematic framework for comparing verification mechanisms, leading to ad hoc decisions in binding agreements like compute pauses or model deployment halts. Concurrently, the AI safety community generally, and technical AI governance researchers specifically, are examining, developing, and testing a wide range of verification methods. The resulting dynamic is one of a fractured gap between policy makers and AI governance advocates, where international policy goals and intent are struggling to keep pace with verification innovations and ideas, and vice versa. We aim to solve this with a product (a decision support engine and dashboard application) that incorporates a Verification Mechanism Feasibility Scorer (VMFS). The VMFS is a multi-dimensional evaluation tool scoring proposals across four dimensions: Evidence Hardness, (the “Trust” score), Infrastructure Burden (The “Cost” Score), Intrusion Burden (the “Friction” score), and Robustness (the “the Cheating Score”). We analyzed nine mechanisms—from on-chip telemetry and TEE attestation to OPCW-style declarations and model watermarking (overall 4.1)—VMFS reveals no silver bullet, suggesting layered approaches will be required in any range of elements of an AI treaty. Our analysis highlights R&D priorities for computer monitoring and consideration of tech transfer to avoid Global South exclusion. Our open-source rubric, mechanism dashboard, and future Global South analysis can empower UN/G7 forums, diplomats, and policy makers to design feasible, halt-capable treaties. Future expanded

¹ Research conducted at the [Technical AI Governance Challenge](#), 2026

elements of the application and dashboard are recommended for improving utility for technical AI government researchers as a beneficiary.

1. Introduction

International AI governance is shifting from principles to enforceable treaties like compute thresholds ($>10^{27}$ FLOPs) and model deployment halts, but verification—credibly confirming compliance amid secrecy—remains unsolved, risking ineffective pacts. The 2023 Bletchley Declaration (UK Gov, 2023), 2024 Seoul Summit (Ministry of Foreign Affairs, Korea, 2024), and 2025 Paris AI Action Summit (ITU Publications et al., 2025) call for "verifiable safety commitments" without tools to compare proposals from (Wasil et al., 2024)(14 methods), Simon Institute IAEA mappings (Krawec, 2025), or Oxford Martin analyses (Harack et al., 2025).

We introduce the Verification Mechanism Feasibility Scorer (VMFS) rubric and dashboard evaluating nine mechanisms (e.g., TEE attestation, remote sensing) across four dimensions. The first is Evidence Hardness, or the "Trust" score, examining whether the evidence is based on math/physics (Hard) or human trust (Soft). Secondly, the Infrastructure Burden, or the "Cost" score, explores whether it requires building new hardware (High Burden) or just passing laws (Low Burden). Next, the Intrusion Burden, or the "Friction" score, covering how much secret data (IP, Weights) is exposed. And lastly, Robustness, or the "Cheating" score, to examine how easy or hard it is to spoof or hide from this mechanism.

Research Questions

1. How can diverse AI verification mechanisms be systematically compared across dimensions relevant to treaty implementation?
2. Which mechanisms are ready for near-term deployment; what are their principal barriers?
3. How can application tools and dashboards engage policy makers, treaty negotiators, and AI governance researchers to support their needs and education?

Our main contributions are:

1. We developed VMFS, the first rubric scoring nine AI verification mechanisms and four feasibility dimensions to prioritize options for halt-capable treaties.
2. VMFS analysis scores mechanisms, revealing trade-offs like model watermarking's 4.1/5 overall lead and TEEs' technical strength amid high infrastructure burden and intrusion, prioritizing compute monitoring R&D.

3. We develop an interactive decision-support system, including a dashboard, portfolio sandbox, and submission interface; enabling users to explore mechanism trade offs, identity coverage gaps, and construct layered verification portfolios.

2. Related Work

AI verification draws from arms-control precedents like IAEA material accountancy (for compute), OPCW declarations (self-reporting), and START national technical means (remote monitoring), as mapped by Simon Institute (Krawec, 2025) and cataloged in (Wasil et al., 2024)(14 methods). Technical proposals include compute governance (Sastry et al., 2024) (Sastry et al., GovAI), hardware telemetry/TEEs (Kulp et al., 2024)(Kulp et al., RAND), model audits/lineage (Harack et al., 2025)(Oxford Martin AIGI), and UN SAB confidential verification. These works propose mechanisms but lack unified multi-dimensional feasibility scoring (e.g. evidence hardness, infrastructure burden, evasion robustness) for treaty prioritization; VMFS fills this gap.

3. Methods

VMFS combines literature review and expert input with a rule-based, text-derived scoring procedure applied to nine verification mechanisms across four feasibility dimensions, producing an interactive dashboard for comparative analysis and portfolio exploration.

4. Results

VMFS scores nine mechanisms (1-5 scale, equal-weighted average) across dimensions, showing variation with no silver bullet above 4.1 overall. Model Watermarking leads at 4.1 (low intrusion). Hardware-anchored mechanisms (e.g., TEE attestation) achieve high evidence hardness (5.0) but are constrained by infrastructure burden (1.0) and intrusion (4.0), resulting in lower scores (3.4).

Key Findings

1. No mechanism tops 4.1 across dimensions; layered portfolios essential (e.g., remote sensing + audits + whistleblowers).
2. Trade-offs are systematic: high-hardness mechanisms impose high burden and intrusion, while low-cost mechanisms sacrifice evidentiary strength.
3. Verification value clusters by evidence location—audits (model safety), telemetry (compute use), sensing (facilities)—with whistleblowers addressing undeclared activity.

5. Discussion and Limitations

VMFS results underscore that AI verification demands pragmatic, layered architectures over idealistic singletons, with implications for near-term treaties prioritizing politically tractable starters like model watermarking(4.1) remote sensing (3.6 avg.) and declarations (OPCW-inspired) to cue deeper checks (e.g., audits for MV7 safety). For AI safety, this enables "halt-capable" regimes verifying compute pauses or unsafe deployments amid competition, reducing surprise risks via confidential evidence (UN SAB), while future work on Global South dimensions could highlight equity imperatives—hardware-heavy methods risk North-South divides without tiered/tech-transfer provisions, mirroring NPT Article IV. Methodological limits include expert-judgment scoring (subjective; needs validation surveys), equal dimension weights (may undervalue politics), and static 2026 snapshot (tech evolves rapidly). Scope excludes novel mechanisms (e.g., ZKPs) or full VFMC depth due to hackathon time; unaddressed threats: insider collusion bypassing whistleblowers (M6) or quantum breaks on TEEs. Differently, we'd weight dimensions dynamically and pilot with negotiators for calibration.

Limitations

VMFS relies on expert judgment for scores rather than empirical data, risking subjectivity—different panels may vary (e.g., political tractability optimistic/pessimistic). Equal dimension weighting ignores priorities (e.g., sovereignty > technical for some treaties); no sensitivity analysis included.

Nine mechanisms illustrative, not exhaustive (e.g., omitted proof-of-learning, model cards); these cover core but not niche risks like multi-agent systems. Scores static—tech/politics evolve rapidly (e.g., TEE maturity by 2027?).

Three-day timeframe limited empirical validation, Global South consultations, and user design/testing of full dashboard refinement (prototype functional, not production).

Future Work

Conduct expert surveys to validate/refine VMFS scores; add sub-dimensions (e.g., hardware vs. software technical) and weighted scenarios (e.g., US-China treaty). This would include deeper user interviews and problem needs identification with technical AI governance researchers seeking solutions for how to improve feedback loops from policy makers and regulators.

Score additional mechanisms (proof-of-learning, model cards, capability evals); extend to multi-agent/emergent risks; integrate VFMC failure modes quantitatively.

Productionize dashboard (web app, API); pilot with users from UN/G7/INAI for treaty design; engage Global South via workshops/tech transfer pilots to iterate adoptability.

Publish regular updates to the “stakeholder solicitation” of verification mechanisms via the VMFS Dashboard, every 3 to 6 months. The evolving compendium of mechanisms would be submitted to

external review panels of AI safety experts and policy-makers, and published as regular version updates for the benefit of the AI safety community, researchers, and other stakeholders.

Identify existing AI safety organizations that would be interested in hosting the VMFS dashboard and related research as part of regular programming, research, and stakeholder engagement/field building.

6. Conclusion

VMFS systematically scores nine AI verification mechanisms, revealing no silver bullet—model watermarking leads overall (4.1 avg) for low intrusion, but layered approaches (e.g., audits + telemetry + whistleblowers) are essential for robust evidence-location coverage in halt-capable treaties. Hardware excels technically (TEE evidence hardness 5.0) are constrained by infrastructure burden (1.0) and intrusion (4.0), motivating prioritization of compute-monitoring R&D and complementary low-friction mechanisms.

The VFMS dashboard empowers negotiators with decision support tools to prioritize feasible regimes, bridging policy-tech gaps for UN/G7/INAI. The open-source rubric enables global adaptation, potentially accelerating intent to generate verifiable commitments for managing frontier risks.

Code and Data

- **Code repository:** [[Link to github](https://github.com/r34l-rudr44/vmfs)] <https://github.com/r34l-rudr44/vmfs>
- **Video Walkthrough:** [[Link to demo video](https://sites.google.com/view/vmfs-demo-video/home)]
<https://sites.google.com/view/vmfs-demo-video/home>
- **Live App:** <https://www.vmfs.app/>

References

Harack, B., Trager, R. F., Reuel, A., & Manheim, D. (2025). Verification for International AI

Governance. *Oxford Martin School, July*(2025), 172.

https://aigi.ox.ac.uk/wp-content/uploads/2025/07/Verification_for_International_AI_Governance.pdf

ITU Publications, AI Governance Initiative, Oxford Martin School, & University of Oxford.

(2025). *The Annual AI Governance Report 2025*. ITU Publications. 978-92-61-41441-2

- Krawec, C. (2025, 09 18). *Mapping IAEA Verification Tools to International AI Governance: A Mechanism-by-Mechanism Analysis*. Simon Institute of Longterm Governance.
<https://simoninstitute.ch/blog/post/mapping-iaea-verification-tools-to-international-ai-governance-a-mechanism-by-mechanism-analysis>
- Kulp, G., Gonzales, D., Smith, E., Heim, L., & Puri, P. (2024, 01 18). *Hardware-Enabled Governance Mechanisms*. RAND.
https://www.rand.org/pubs/working_papers/WRA3056-1.html
- Ministry of Foreign Affairs, Korea. (2024, 05 23). *SEOUL DECLARATION*. Ministry of Foreign Affairs. https://www.mofa.go.kr/eng/brd/m_5674/view.do?page=1&seq=321007
- Sastry, G., Heim, L., Belfield, H., Anderljung, M., & Brundage, M. (2024, 02 13). *Computing Power and the Governance of Artificial Intelligence*. arXiv.
<https://arxiv.org/html/2402.08797v1>
- UK Gov. (2023, November 2). The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023. *GOV.UK*. Retrieved February 2, 2026, from <https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023>
- Wasil, A. R., Reed, T., Miller, J. W., & Barnett, P. (2024, 11 4). *Verification methods for international AI agreements*. arXiv. <https://arxiv.org/abs/2408.16074v2>

Appendices

Appendix A: VMFS Scoring Framework

A.1 Dimension Definitions:

The VMFS evaluates each verification mechanism (M1–M9) across four dimensions, with scores derived strictly from mechanism characteristics documented in the database.

Dimension	Anchor Question	Source Field
Evidence Hardness ("Trust")	Is the evidence based on math/physics (Hard) or human trust (Soft)?	"Evidence it produces"
Infrastructure Burden ("Cost")	Does it require building new hardware (High Burden) or just passing laws (Low Burden)?	"What it needs to work"
Intrusion Level ("Friction")	How much secret data (IP, Weights) is exposed?	"Biggest uncertainty/limitation"
Evasion Robustness ("Cheating")	How easy is it to spoof or hide from this mechanism?	"Biggest uncertainty/limitation"

A.2 Scoring Decision Trees

Evidence Hardness (1.0–5.0)

- **5.0 (Immutable/Physical):** Cryptographic signatures, signed hashes, telemetry, sensor readings, tamper-evident logs
- **3.0 (Digital/Mutable):** Unsigned databases, removable watermarks, API responses
- **1.0 (Subjective/Human):** Self-reports, declarations, testimony, whistleblower accounts

Infrastructure Burden (1.0–5.0)

- **5.0 (Policy/Software Only):** Laws, regulations, public registries, whistleblower protections
- **3.0 (Commercial/Cloud):** Vendor cooperation, API integration, commercial imagery

- **1.0 (New Frontier Hardware):** Specialized chips, TEEs, satellite constellations, on-site inspection teams

Intrusion Level (1.0–5.0)

- **5.0 (Zero/External):** Aggregate data, thermal signatures, remote observation
- **3.0 (Metadata/API):** Traffic analysis, query access, header inspection
- **1.0 (Deep/Internal):** Weights, training data, IP exposure, always-on surveillance

Evasion Robustness (1.0–5.0)

- **5.0 (Airtight):** No viable evasion path; mathematically or physically undeniable
- **3.0 (Costly Evasion):** Bypass possible but requires significant effort
- **1.0 (Known Gaps):** Explicit coverage gaps, legacy hardware vulnerabilities, suppression risks

Appendix B: Evidence Location Architecture

B.1 Verification Goals and Evidence Mapping

The dashboard translates policy intent into technical architecture via a relational mapping:

Goal → Evidence Location → Mechanism.

Verification Goal	Evidence Type	Evidence Location
Geographical location of compute clusters	GPS-stamped heartbeats, thermal signatures	The Data Center
Total FLOPs during training	Cryptographically signed GPU telemetry	The Chip
Hardware origin legitimacy	Supply-chain manifests, serial registries	The Hardware Supply Chain
Training data filtering compliance	Audit certificates, provenance hashes	The Developer Lab

Dataset licensing basis	Digital contracts, usage-right metadata	The Institutional Record
Model identity and version history	Weight-hashes tied to authorized registry	The Central Model Registry
Post-release modifications	Signed change logs, re-hashed weights	The Central Model Registry
Access control enforcement	Real-time access logs, identity tokens	The Model Deployment Point
Hardware-level safeguards	TEE attestations, tamper-evident seals	The Chip
Authorized server environment	Server identity attestations, IP verification	The Model Deployment Point

B2. Mechanism Library

ID	Mechanism	Evidence Location	Proof Provided
M1	On-Chip Telemetry	The Chip	Signed logs of total FLOPs and runtime
M2	Chip Registry	The Hardware Supply Chain	Ledger of serial numbers and transfer-of-custody
M3	TEE Attestation	The Chip	Signed proof of software integrity and secure execution
M4	Third-Party Audits	The Developer Lab	Capability risk reports and safety-compliance certificates
M5	Model Watermarking	The Model Deployment Point	Detectable signals in output proving model origin

M6	Whistleblower Programs	The Institutional Record	Insider testimony and internal documents
M7	Remote Sensing	The Data Center	Satellite imagery showing facility scale and activity
M8	Declaration Regimes	The Institutional Record	Formal self-reports on infrastructure and safety tests
M9	Data Filtering	The Developer Lab	Logs showing removal of prohibited content hashes

Appendix C: Dashboard Components

C.1 Portfolio Sandbox

The Portfolio Sandbox enables users to simulate treaty regimes by combining 1–3 mechanisms.

Scoring Logic:

- Compute the average across four dimensions for the selected mechanism set
- Visualize via radar chart (North=Hardness, South=Low Burden, East=Robustness, West=Low Intrusion)

Gap Analysis (Red Team Alert): The system identifies blind spots in user-selected portfolios.

Example logic:

IF selected_locations NOT INCLUDE chip_level THEN

Display: "Your portfolio cannot verify digital activity (FLOPs).

A bad actor could use legal hardware for illegal training.

Consider adding On-Chip Telemetry."

C.2 Verification Advisor

The Advisor is an LLM-based chatbot serving as a Technical Auditor persona, prioritizing "Proof" and "Cost" over diplomatic framing.

Operational Rules:

1. Always cite specific scores (e.g., "Remote Sensing has Robustness score: 3.0")
2. Emphasize trade-offs (e.g., "High security but high Infrastructure Burden")
3. Reference Evidence Locations to explain mechanism coverage

Example Prompts:

- "Find me a high-trust (Hardness > 4) mechanism."
- "What is the cheapest way to verify compute?"
- "Why does On-Chip Telemetry have a high Intrusion score?"

C.3 Stakeholder Submission Interface

Users can submit novel verification ideas for evaluation against the VMFS framework.

Extraction Pipeline:

1. **Evidence Location Identification:** Map user input to relevant location (e.g., "registry of researchers" $\rightarrow$ Institutional Record)
2. **Score Prediction:** Apply scoring rubrics to generate provisional dimension scores
3. **Output Generation:** Create temporary Mechanism Card with radar chart visualization

Example:

- Input: "A global registry of AI researchers"
- Predicted Scores: Hardness 1.0, Burden 5.0, Intrusion 2.0, Robustness 2.0
- Interpretation: "Low-cost but low-trust mechanism with privacy concerns"

Appendix D: Limitations & Dual-Use Considerations

Please see limitations as listed in the main section of the paper.

Dual-use risks involve the potential for malicious actors or AI systems to use this information to design or anticipate ways to avoid verification mechanisms, complicate their negotiation, or otherwise undermine coordination efforts within the AI safety and political realms on international agreement utilizing verification and enforcement mechanisms.

Appendix E: Responsible Use Declaration

VMFS is a decision-support tool designed to inform—not replace—expert judgment in treaty negotiations. Users should recognize that:

1. **Scores are indicative, not prescriptive.** Feasibility scores reflect technical and institutional characteristics under idealized conditions; real-world deployment requires context-specific legal, diplomatic, and political analysis beyond this framework's scope.
2. **The tool does not endorse mechanisms.** High scores indicate implementation feasibility, not normative desirability. Ethical, sovereignty, and human rights implications require separate assessment.
3. **Intended users are policy professionals.** VMFS assumes users possess baseline familiarity with international governance; misapplication by untrained actors may yield misleading conclusions.
4. **Open-source release supports global adaptation.** We encourage modification for regional contexts while requesting attribution and responsible deployment aligned with cooperative AI governance goals.

Appendix F: Ethical Considerations

- **Privacy and Sovereignty.** Several high-scoring mechanisms (e.g., on-chip telemetry, TEE attestation) require intrusive data access. We flag Intrusion scores explicitly to ensure negotiators weigh surveillance risks against verification benefits, particularly for jurisdictions with strong data protection norms.

- **Equity in Verification Burden.** Hardware-dependent mechanisms risk imposing disproportionate compliance costs on Global South nations lacking domestic semiconductor industries. We highlight this asymmetry qualitatively and recommend future work on tiered implementation and technology transfer frameworks.
- **Transparency-Security Balance.** Full methodological disclosure enables scrutiny and reproducibility but may reduce operational security if deployed in sensitive negotiations. Users should adapt transparency levels to the appropriate context.

LLM Usage Statement

We used Claude, Gemini AI, to assist with brainstorming, research, interface copy, and rapid prototyping during dashboard development. All design decisions, claims, and results were reviewed and independently verified by the authors.

January 2026