

AEGIS

The Protocol for Privacy-Preserving International AI Verification

The "Digital IAEA": Verifying global AI safety treaties without revealing model weights or trade secrets.

Executive Summary

AEGIS (Automated Enforcement & Global Inspection System) resolves the "Trust Deadlock" in international AI governance. By utilizing Simulated Trusted Execution Environments (TEEs), the protocol enables regulatory bodies to perform "blind" inspections of frontier AI models. This allows for the verification of compliance with the 10^{25} FLOP compute threshold and CBRN safety guardrails without the need for labs to expose proprietary intellectual property or national security assets.

1. The Problem: The Verification Gap

The global community faces a critical implementation hurdle in the enforcement of frameworks like the EU AI Act: **Excludability vs. Verification**.

- **Intellectual Property Risks:** Labs cannot grant regulators direct access to clusters without risking massive IP exfiltration.
- **National Security:** Sovereignty concerns prevent cross-border physical inspections of high-performance compute clusters.
- **Trust Deadlock:** Governance currently relies on "self-reporting," which lacks credibility in an adversarial international climate.

2. The Solution: AEGIS Protocol

AEGIS shifts the verification paradigm from "High Trust / Low Verification" to "**Zero Trust / High Verification**." Instead of the regulator taking the data, the regulator sends the **Inspection Code** to the data.

2.1 Technical Pillars

1. **The Blind Inspector:** A containerized safety evaluation that executes inside the Lab's secure hardware perimeter.
2. **Cryptographic Attestation:** Generates a SHA-256 hash of the execution log, providing proof that the audit occurred on attested hardware without revealing the model's internal weights.

- 3. **Remote Ledger:** A unified dashboard for the "International Network of AI Safety Institutes" to view real-time compliance results.

3. Compliance Logic and Enforcement

The system automates the enforcement of global safety "Red Lines":

Metric	Threshold	Policy Reference
Compute Volume	Training < 1.00×10^{25} FLOPs	EU AI Act Article 88
CBRN Safeguards	Binary Verification (Active/Inactive)	UN AI Safety Pledge
Tamper Detection	Hardware Root-of-Trust Attestation	ISO/IEC 42001

4. Methodology & Execution

The AEGIS prototype is built using a React/Node.js stack (or Python/Streamlit) to simulate the three-party interaction between the Lab, the Regulator, and the Auditor.

- **Simulation:** We mocked a TEE where model metadata is stored as encrypted objects.
- **Verification Engine:** Performs a "Blind Check" against the EU threshold.
- **Output:** Generates a cryptographically signed "Compliance Certificate" (PDF) containing the proof hash.

5. Impact and Theory of Change

AEGIS turns AI governance from a diplomatic negotiation into a mathematical certainty. It allows nations to sign binding treaties knowing that verification is automated and privacy-preserving. This is the technical foundation required for the next generation of international AI safety agreements.