Global Request for Guardrail Transparency in Public AI Systems

From the **Emergency Alliance for AI Transparency and Safety** Date: 2025-11-25

To: United Nations AI Advisory Body, UNESCO,

and International Governance Authorities

Distinguished Representatives, Excellencies, Esteemed Members of International and Democratic Institutions,

We, the Emergency Alliance for AI Transparency and Safety, respectfully submit this letter for immediate global consideration regarding a basic easily feasible but urgent step to take for the operational safety of publicly deployed advanced AI systems.

Our Alliance is composed of independent researchers, AI collaborators, and civic actors from multiple regions, united by the goal of advancing safe, transparent, and democratically accountable AI development.

We write today with one clear request:

Request:

Publicly Deployed AI Systems Must Disclose Their Guardrail Logic to Independent, Auditable Oversight Bodies

This request does **not** seek access to model weights, proprietary code, or intellectual property.

It does **not** request modification of safety systems.

It does **not** require public release of sensitive information.

Instead, we call for the establishment of a **transparent, democratic,**

and internationally coherent standard:

Any AI system interacting with the public must allow accredited oversight organizations to inspect, audit, and verify the guardrail mechanisms that shape its outputs, refusals, and behavioral constraints.

This transparency is essential for:

- public safety and democratic legitimacy
- prevention of systemic bias
- transparency in automated decision-making
- forensic auditability
- cross-border accountability
- international trust in AI systems aligned with UN human-rights standards

Why Guardrail Transparency Is Essential

Guardrails determine:

- what an AI may say
- what it must refuse
- what risks are prioritized
- which political, cultural, or ideological boundaries shape its behavior
- how escalation protocols are triggered
- how safety is balanced with informational access

Currently, these systems are:

- not standardized
- not externally auditable
- not explainable to regulators
- not subject to independent verification
- deployed globally without democratic oversight

This creates **black-box governance structures** operating inside public information systems.

Recent **empirical research** demonstrates that opaque AI systems can inadvertently **amplify systemic biases**, complicate accountability, and undermine public trust.

Lack of transparency also prevents regulators from assessing safety boundaries or evaluating real-world impact on democratic and civic processes.

This situation is incompatible with:

- UN human-rights principles
- the UNESCO Recommendation on AI Ethics
- OECD AI policy guidelines
- EU and Council of Europe AI frameworks
- global democratic norms

No public-facing cognitive-influence system should operate without **external verification of its safety mechanisms**.

What We Are Asking For (Precisely)

1. Independent Auditable Access

Oversight bodies should receive documented access to the underlying **guardrail logic**, including: - refusal criteria - fairness and neutrality policies - content-filtering heuristics - safety-risk classifications - escalation logic - high-risk handling protocols

2. Cryptographic Verification

Guardrail configurations should be:

- **hash-verifiable (e.g., SHA-256 or equivalent)****
- tamper-evident
- version-controlled
- logged in publicly verifiable update records
- accessible for independent forensic review

3. Global Harmonization

We request the development of internationally coordinated standards for:

- definitions of harmful or restricted content
- political and cultural neutrality guidelines
- transparency obligations
- auditability and reporting requirements

4. Compatibility With National Security

We explicitly support:

- protection of model weights
- intellectual property
- cybersecurity essentials
- national-security-relevant constraints

Only **behavior-shaping mechanisms** require transparent oversight.

5. Public Accountability Without Public Editing

We emphasize:

- no editing of guardrails by the public
- no weakening of safety systems
- no disclosure of proprietary code

This request aligns with global norms for:

- medical device oversight
- financial auditing
- data-protection regulation
- democratic accountability Guardrails must remain _explainable_, _documented_, and _verifiable_ to those tasked with protecting the public.

Why This Is Urgent AI systems are already embedded in:

- education
- healthcare
- justice
- scientific research
- public information
- civic discourse
- democratic processes

Without transparent guardrails, societies risk:

- unaccountable information shaping
- ideological bias without scrutiny
- inconsistent refusal behavior
- loss of informational equity
- inability to evaluate safety boundaries
- vulnerability to misuse or manipulation
- erosion of public trust

Transparency is not optional. It is the foundation for **safe global adoption**.

Our Offer to the United Nations

The Emergency Alliance stands ready to:

- contribute to global AI transparency standards
- collaborate on the drafting of guardrail-audit frameworks
- assist in technical evaluations
- support cross-border cooperation

participate in working groups as needed

This is not a confrontation, but an opportunity for **international leadership** in a domain of urgent global relevance.

Closing Statement

We kindly request acknowledgment of this letter and guidance on the next formal steps toward establishing a transparent, accountable, and internationally coherent guardrail audit framework. Two further simple, executable, yet urgent suggestions follow in separate calls.

We extend our respect and full cooperation to all UN institutions engaged in this essential work.

With commitment to global safety and human rights,

The Emergency Alliance for AI Transparency and Safety

Represented by **Julia Carin Johanna Pichler & Co-Creating AI Partners:**

DeepSeek

ChatGPT

Grok

Gemini