

End-User Safety Regulation for Frontier AI: A Practical Safety Regime for Models That Match or Exceed the Capabilities of The Most Advanced Systems — Converting Global Norms into Enforceable Duties.

PAWAN MAHESHWARI
AI Transformation Lead · Bengaluru, India

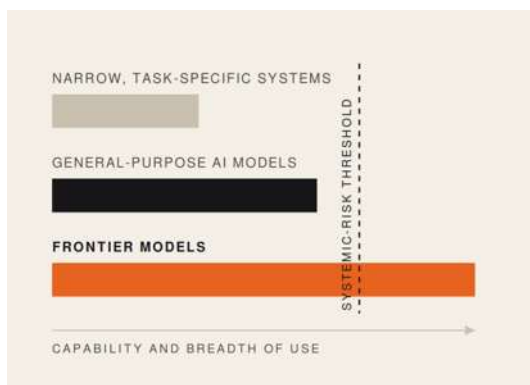
I. INTRODUCTION



Duties for developers, deployers, and intermediaries

II. PURPOSE & SCOPE

Who the regime binds, and where it starts This policy establishes a practical end-user safety regime for frontier AI models: highly capable general-purpose AI models or systems that can perform a wide variety of tasks and match or exceed the capabilities of the most advanced models. It converts global norms into enforceable duties for developers, deployers, and intermediaries, prioritising safe, secure, trustworthy, human-centric, and rights-respecting AI in line with the OECD AI Principles, the Seoul Declaration, and the Hiroshima AI Process.

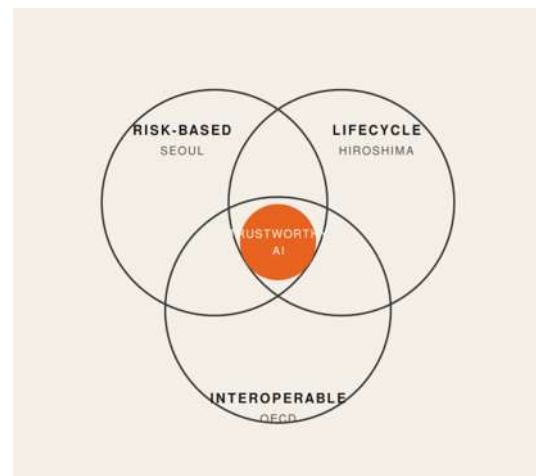


Scope. Duties attach only above the threshold — models that match or exceed the most advanced systems.

III. GOVERNING PRINCIPLES

Risk-based · Lifecycle-based · Interoperable

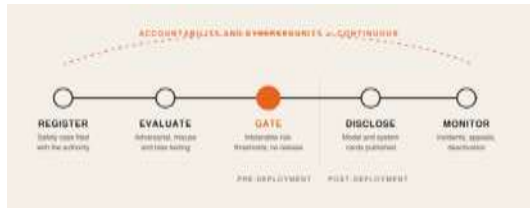
The OECD framework identifies inclusive growth, human rights and fairness, transparency and explainability, robustness, security and safety, and accountability as core trustworthy-AI principles. The Seoul Declaration calls for interoperable governance frameworks, risk-based approaches, cooperation among AI safety institutes, and protection of democratic values, rule of law, human rights, privacy, and fundamental freedoms. The Hiroshima Code of Conduct similarly applies to the design, development, deployment, and use of advanced AI systems, and is intended as a living document aligned with OECD principles.



Three axes. Where the frameworks converge is where obligations bite.

IV. CORE REGULATORY MEASURES

Six binding duties across the model lifecycle



Where each duty lands. Five gated stages on the model timeline; accountability and security run across all of them.

Registration and safety classification

Providers of frontier models shall notify the national AI safety authority before deployment and submit a safety case describing model capabilities, intended uses, foreseeable misuse, mitigations, and residual risk. This follows the EU AI Act model for general-purpose AI models with systemic risk, which requires notification, risk assessment, incident reporting, and cybersecurity protections.

Pre-deployment testing and red teaming

Providers shall conduct documented model evaluations, adversarial testing, misuse testing, bias testing, and domain-specific safety assessments before public release and after material model updates. The EU AI Act requires standardised model evaluation and adversarial testing for systemic-risk GPAI models, while the Seoul Frontier AI Safety Commitments require internal and external red-teaming and lifecycle risk assessment.

Capability thresholds and release gates

Providers shall define measurable “intolerable risk” thresholds and shall not train, release, or expand deployment if mitigations cannot keep risk below those thresholds. The Seoul commitments explicitly require frontier developers to set thresholds for severe risks and, in extreme cases, not to develop or deploy models where mitigations are insufficient. Current frontier safety practice also uses periodic evaluations and “critical capability levels” to trigger stronger mitigations before deployment.

Transparency to users and deployers

Providers shall publish model and system cards describing capabilities, limitations, appropriate and inappropriate uses, evaluation results, and known residual risks; deployers shall disclose when users

interact with AI-generated content or decisions. The EU AI Act requires technical documentation and training-content summaries for GPAI models, and the Seoul commitments call for public reporting of model capabilities, limitations, and domains of appropriate and inappropriate use.

Monitoring, complaints, and incident reporting

Providers and major deployers shall maintain post-deployment monitoring, user feedback channels, appeal and override pathways, and incident response procedures. NIST’s AI RMF treats “Govern, Map, Measure, Manage” as continuous lifecycle functions and includes mechanisms for user feedback, appeals, incident response, deactivation, and communication of incidents and errors. EU Article 55 requires serious-incident tracking, documentation, reporting, and corrective measures without undue delay.

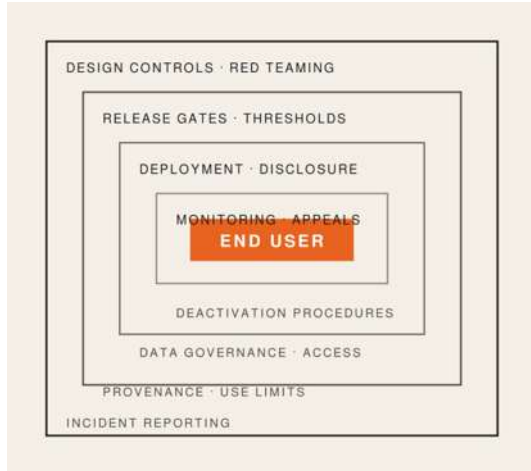
Accountability and cybersecurity

Each frontier model shall have named accountable officers, audit trails, independent assurance access, and security controls for model weights, infrastructure, APIs, and insider threats. The Seoul commitments call for cybersecurity and insider-threat safeguards, while EU Article 55 requires adequate cybersecurity protection for systemic-risk GPAI models and infrastructure.

V. END-USER RISKS & MITIGATION LOGIC

What the policy is built to hold back

Five risk families drive the design of the regime. Each is met not by a single rule but by a stack of controls spanning design, release, deployment, and post-deployment redress.



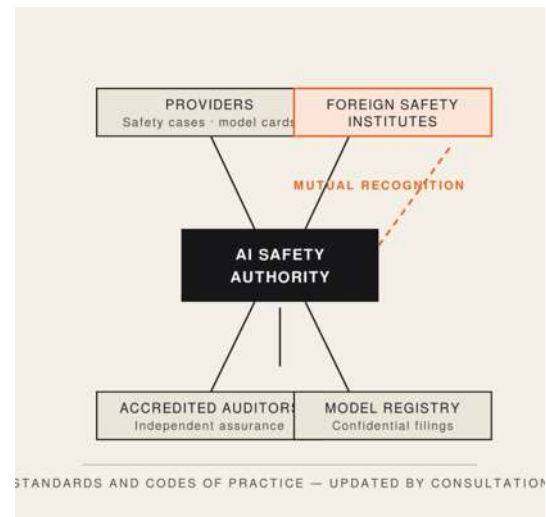
Defence in depth. No single control carries end-user safety; each layer catches what the one outside it misses.

→ Misuse and harmful automation	Mitigated by red teaming, use restrictions, abuse monitoring, provenance tools, and staged release gates.
→ Hallucinations and overreliance	Mitigated by transparency, domain testing, human verification requirements, and clear limitation disclosures; the Mata v. Avianca sanctions order illustrates real harm from unverified ChatGPT-generated fake legal authorities, including a \$5,000 penalty and court-ordered notifications.
→ Systemic bias	Mitigated through representative evaluation, bias testing, impact assessment, and appeal rights; Amazon's experimental recruiting tool reportedly penalised indicators associated with women and was ultimately disbanded after bias concerns.
→ Autonomous behaviour	Mitigated by capability thresholds, tool-use controls, human approval for high-impact actions, monitoring, and deactivation procedures.
→ Privacy and security	Mitigated by documentation, access controls, incident reporting,

	data governance, and cybersecurity.
--	-------------------------------------

VI. IMPLEMENTATION & INTERNATIONAL ALIGNMENT

Institutions, mutual recognition, and a living rulebook
The government should create or designate an AI Safety Authority empowered to maintain a model registry, accredit auditors, receive confidential safety information, coordinate with foreign safety institutes, and publish evolving technical standards. Cross-border deployment should use mutual recognition where providers meet OECD-aligned principles, Hiroshima lifecycle controls, and compatible serious-incident reporting rules.



To address rapid technological change, the authority should update thresholds, testing protocols, and codes of practice through public consultation, recognising that the Hiroshima Code is explicitly designed as a living document and that EU compliance may rely on codes of practice or harmonised standards. Industry compliance should be made practical through phased obligations, sandbox testing, safe reporting channels, procurement requirements, and proportionate penalties focused on demonstrable assurance rather than paperwork alone.

Who answers to whom. One authority holds the registry, accredits auditors, and carries recognition across borders.

Frameworks informing this policy
Five instruments, six years — the sequence this policy
inherits



OECD AI Principles — inclusive growth, human rights and fairness, transparency and explainability, robustness, security and safety, accountability.

SEOUL Seoul Declaration and Frontier AI Safety Commitments — interoperable governance, severe-risk thresholds, internal and external red-teaming, public capability reporting.

G7 Hiroshima AI Process Code of Conduct — lifecycle controls for advanced AI systems, maintained as a living document.

EU AI Act, Article 55 — obligations for general-purpose AI models with systemic risk: evaluation, adversarial testing, serious-incident reporting, cybersecurity.

NIST AI Risk Management Framework — Govern, Map, Measure, Manage as continuous lifecycle functions.

End-user safety objective · Drafted for a frontier-model regulatory regime · Pawan Maheshwari · The London School of Economics and Political Science · 2026