POLICY

PAPER

SOVEREIGNTY, ACCESS, AND EXTREME RISKS:

A Three-Layer Architecture for International AI Governance

Daniel Stauffacher

GENEVA 2026
ICT4Peace Foundation

This paper is published by the ICT4Peace Foundation. Comments are welcome and may be addressed to the Foundation.

ICT4Peace Publishing, Geneva, July 2026
Copies available from www.ict4peace.org

The text was prepared with the assistance of Claude (Anthropic).

Executive Summary

Does the rise of "Sovereign AI" undermine international coordination? The concern is widely voiced. The UN Secretary-General has warned against "fragmented AI spheres"[1]; more than eighty countries are advancing AI strategies built on divergent regulatory philosophies[2]; and 118 countries, most of them in the Global South, are party to none of the major international AI governance initiatives[3]. This paper argues that the sovereignty trend, properly understood, can serve a constructive purpose: it forces a distinction between what genuinely requires a binding global agreement and what does not. Our model is the response to the 2008 financial crisis, when the world agreed on shared standards only for banks whose failure would endanger the global system, and left the rest of banking regulation national[4]. We propose the same structure for AI:

1. A **narrow binding core** for extreme, inherently transboundary risks — loss of control over frontier systems, and the misuse of advanced AI for biological, cyber, or military harm — applying only to systems above a compute-and-capability threshold.

2. A **thicker layer of voluntary technical cooperation**: shared evaluation methods, common incident-reporting formats, model documentation standards.

3. **Full national discretion** everywhere else: deployment rules, content governance, data policy, industrial strategy.

No binding regime on extreme risks will attract broad support unless access is part of the same bargain. Frontier model development is concentrated in a handful of jurisdictions, so access commitments (helping countries adapt open models, pooling regional compute, treating open foundation models as public goods) must sit alongside risk obligations, much as the nuclear non-proliferation regime paired restrictions with a promise to share peaceful technology. Switzerland's fully open Apertus model shows what such a public good can look like at national scale.

Finally, the paper treats the AI divide as a development question. For the Global South the obstacles compound one another, from energy, connectivity and compute to data, talent and fiscal space. Development cooperation should therefore treat AI as core development policy: integrated packages rather than one-off projects, compute access as infrastructure, language and data resources as public goods, and a funded seat at the table for developing countries in global governance processes.

[1] António Guterres, UN Secretary-General, remarks to the United Nations Security Council on artificial intelligence and the maintenance of international peace and security, New York, 19 December 2024.

[2] T. Oweke, *The World Is Trying to Govern AI. The UN Wants In.*, Council on Foreign Relations, May 2026.

[3] United Nations High-level Advisory Body on Artificial Intelligence, *Governing AI for Humanity* — Final Report, September 2024.

[4] Basel Committee on Banking Supervision / Financial Stability Board, framework for Global Systemically Important Banks (G-SIBs), 2011, as updated. See Section 3 and reference 10.

1. The Question

Sovereign AI is one of the defining trends of the mid-2020s. States are pouring money into national compute infrastructure, domestically trained models, and independent regulatory regimes. Does this make international coordination harder? Many observers fear it does: the outlines of competing technology blocs (export controls on advanced chips, restrictions on cross-border data flows, divergent regulatory regimes, rival standards) are visibly hardening, and the UN Secretary-General has repeatedly warned that a world of "fragmented AI spheres" would be one of perpetual instability.[5] So is there an institutional balance between national sovereignty and the need for global governance, one that avoids fragmentation without overriding each country's own development path?

The tension is real, but it is usually misdiagnosed. The difficulty lies less in sovereignty itself than in *undifferentiated* coordination. The recurring weakness of current AI governance efforts is the attempt to govern "AI" as a whole, instead of the narrow set of points where national systems actually collide.

2. Disaggregating "Sovereign AI"

"Sovereign AI" bundles together three separate claims, each with a different relationship to international coordination:

- **Infrastructure sovereignty**: national or regional control over compute, data centres, and cloud services. This is mainly a resilience agenda, and it is largely compatible with international coordination.
- **Capability sovereignty**: the ambition to train frontier-class models domestically. This is what generates competitive dynamics that can erode a country's incentive to accept external safety constraints.
- **Regulatory sovereignty**: the right to set national rules on deployment, data, and content. This is the most direct source of regulatory fragmentation.

Treating these three claims as a single bloc leads to poor policy. Most of the fragmentation risk comes from the second and third categories; the first can proceed alongside, and even in support of, international cooperation. The choice, in other words, is not between sovereignty and cooperation but between two forms of sovereignty: one that divides the world into blocs, and one that connects it. Openness is what makes sovereignty shareable.

One caveat has moved from theory into practice in 2026, though. The architecture proposed here, like most current approaches, including compute-threshold regulation, assumes, implicitly, that AI infrastructure stays civilian in character, identifiable, and stable. Recent events complicate that assumption. ICT4Peace's *Bombing Clouds* analysis documents how commercial cloud infrastructure has been drawn into military operations, pulling hyperscale data centres into the battlespace and blurring the civilian-military line that both international humanitarian law and infrastructure-based governance depend on. Meanwhile, the drift of capable models toward edge deployment is eroding the centralized chokepoints that threshold-

[5] Guterres, remarks to the UN Security Council, 19 December 2024 (see note 1): "A world of AI haves and have-nots would be a world of perpetual instability."

based oversight assumes exist. Neither trend invalidates the architecture below, but both make the case for urgency: the window in which frontier AI stays concentrated, centralized, and governable through identifiable infrastructure may not stay open long.

3. The Precedent: Governing the Systemic Tier

International civil aviation is the analogy most often reached for in AI governance debates. Post-crisis financial regulation is a closer structural match.

Every state keeps sovereign banking supervision, sovereign monetary policy, and sovereign financial rules. Yet after 2008, the international community, through the Basel Committee and the Financial Stability Board, agreed on one narrow thing: banks whose failure would endanger the global system ("global systemically important banks") have to meet shared standards, including extra capital requirements, tighter supervision, and resolution planning. The world did not harmonize finance. It identified the systemic tier and governed just that. Sovereignty and coordination grew together, rather than trading off against each other.

The analogy carries a further lesson for AI, and a precise one: systemically important banks are identified through quantitative thresholds combined with qualitative judgment: size, plus interconnectedness, substitutability, and complexity. The equivalent for AI is a compute threshold paired with demonstrated capability.

4. A Three-Layer Architecture

Layer 1: A Narrow Binding Core for Extreme Risks

The binding core should cover only risks that are inherently transboundary and potentially catastrophic. The Strategic Foresight Group's report *The Essential Convergence* and the Global Compact on Extreme AI Risks, launched in Geneva on 6 July 2026 by ICT4Peace, SFG, IFRC, and GCSP, identify that narrow set: loss of human control over the most advanced systems, and misuse of advanced AI for biological, cyber, or military harm.

Scope definition. A compute-based threshold, in the range of 10^{26}–10^{27} FLOPs of training compute paired with demonstrated capability well beyond the prior state of the art, is a workable international anchor. That range sits one to two orders of magnitude above the EU AI Act's systemic-risk presumption (10^{25} FLOPs), and at or above the threshold used in the since-rescinded 2023 US Executive Order (10^{26} FLOPs). It would capture only the very top of the frontier, and deliberately leave out the broader tier of general-purpose models. The threshold needs to be:

- **reviewed periodically**, since fixed compute numbers are a weakening proxy for capability as algorithmic efficiency improves; and
- **tightly scoped to the identified extreme risks**, so that it does not become a vehicle for broader regulatory ambitions.

The sovereignty argument. No country's development path is meaningfully constrained by agreeing to avoid outcomes that would endanger every development path. The threshold operates as a narrow demarcation line, not as a constitution for the technology as a whole. It also embodies a principle of neutrality: safety rules should follow what systems can do, not who

built them. Obligations defined by capability apply identically to American, Chinese, European and open models, which is precisely what makes them negotiable between rivals.

Layer 2: Technical Interoperability

Below the binding core sits a thick layer of voluntary, mutually reinforcing technical cooperation: shared evaluation and red-teaming methods, common incident-reporting formats, model documentation standards, and mutual recognition of conformity assessments. States and developers have a practical reason to adopt them: divergence is expensive. Most of the practical work of preventing fragmentation takes place at this level, in technical fora, without any formal transfer of sovereignty. The network of AI Safety/Security Institutes, the OECD, and the ITU's standardization apparatus are the natural homes for this layer.

Layer 3: National Discretion

Above these two layers, states keep full discretion: deployment rules, content governance, data localization, procurement, industrial strategy remain sovereign choices reflecting each society's own values and circumstances. Regulatory diversity at this layer should not be mistaken for fragmentation. It is closer to federalism, and it generates useful comparative evidence about which governance approaches actually work.

5. The Access Problem — and Why It Decides Everything

The architecture above answers the coordination question. A harder question, however, lies underneath it: today, only a handful of actors, concentrated in the United States, China, and Europe, hold the full stack needed to develop frontier models: energy, compute, data, talent, and capital. For most states, the pressing issue is not how to govern this technology, but how to obtain access to it in the first place.

Access is not a side issue for extreme-risk governance: without a credible answer to it, no binding regime will materialise. No binding regime on extreme risks will be accepted by most nations if it looks like the countries that already have the technology are pulling up the ladder behind them.

The nuclear non-proliferation regime offers the relevant bargain structure: risk obligations traded for a commitment to share peaceful technology (NPT Article IV). It also offers the relevant warning: that bargain is widely seen in the Global South as under-delivered, and any AI equivalent will only be credible if access is real, funded, and early, rather than a promissory note.

Three concrete access pathways:

1. **Adaptation over duplication.** Meaningful AI sovereignty does not require training frontier models from scratch. Fine-tuning and adapting open models for national languages, legal systems, agriculture, and health delivers most of the developmental value at a fraction of the cost. International support should prioritize this tier.
2. **Regional compute pooling.** No single low- or middle-income country can fund a frontier training run on its own; regional facilities change the arithmetic — much as

CERN did for European physics, and as national public-compute programmes such as the IndiaAI Mission already demonstrate at country scale.

3. **Open foundation models as global public goods.** Nor is openness the idiosyncrasy of one small state: China releases several of its leading models with open weights, which any country can use and adapt, and Switzerland's Apertus model, developed by EPFL and ETH Zurich with fully open weights, data, and training methodology, shows that a national investment can become infrastructure any country can build on. When both an AI superpower and a small neutral state practise openness, sovereignty through openness stops being a theory and becomes a working access channel across bloc lines. Development banks and multilateral funds should treat building and maintaining such open models as fundable public infrastructure, alongside proposals for shared international research facilities (a "CERN for AI").

6. The Development Dimension: Challenges and a Development Cooperation Response

The access pathways above address the technology gap. But for most of the Global South, the AI divide sits on top of older, unresolved development challenges, and will widen them further if development cooperation does not adapt.

6.1 The compounding challenges

The full-stack deficit. Frontier AI needs a stack of interdependent capabilities (reliable energy, compute infrastructure, data, skilled talent, capital), and most developing countries face shortfalls at every layer at once. Isolated interventions therefore fail: donated hardware achieves nothing without power, connectivity, and engineers to run it.

Energy and infrastructure. Data centres need gigawatt-scale, uninterrupted power, in regions where hundreds of millions of people still lack basic electricity. AI infrastructure investment risks competing with basic electrification rather than complementing it, and raises water and grid-stress concerns in climate-vulnerable regions.

Connectivity. Roughly a third of humanity is still offline, concentrated in the least developed countries. AI-enabled services mean little where the underlying connectivity layer — the unfinished business of the WSIS process, is simply not there.

Data and language exclusion. Foundation models are trained overwhelmingly on English, Chinese, and a handful of other high-resource languages. Thousands of languages spoken across Africa, Asia, and the Pacific are barely represented, so model quality, and the economic value that follows from it, ends up lower exactly where development needs are greatest. Data about developing-country contexts (agriculture, health, informal economies) tends to be scarce, unstructured, or held abroad.

Talent and brain drain. AI talent is heavily concentrated in a handful of hubs, and salary differences pull the strongest researchers and engineers out of developing countries. Local universities often lack the compute to teach modern methods at all, which reinforces the cycle.

Fiscal space. Many developing countries are already in debt distress, so large-scale public investment in AI infrastructure competes directly with health, education, and climate

adaptation budgets. Without concessional finance, "national AI strategies" remain on paper rather than becoming capabilities.

Labour-market exposure. Economies built on services exports (business-process outsourcing, call centres, transcription, junior software work) are disproportionately exposed to exactly the tasks current AI systems automate first. The traditional development ladder of moving up through services may get pulled away before many countries have finished climbing it.

Governance capacity. Most developing countries lack dedicated AI regulatory institutions, and many lack even baseline data protection frameworks. They are thus doubly disadvantaged: excluded from developing the technology, and under-resourced to take part in the standard-setting processes, in Geneva and elsewhere, that will end up governing it.

Dependence as exposure. One challenge only became concrete in 2026: reliance on foreign commercial cloud infrastructure exposes the civilian life of dependent countries to armed conflicts they have no part in. When hyperscale data centres in the Gulf were struck in the June 2026 hostilities, on the grounds that their computing capacity supported one belligerent's military operations, the resulting outages cascaded through banking, payments, transport, and logistics in neighbouring states with no role in the conflict (see ICT4Peace, *Bombing Clouds*, 2026). For countries whose government services, financial systems, and health records run on infrastructure they neither own nor host, digital dependence is no longer only an economic vulnerability; it has become a civilian-protection concern. That gives the access agenda in this paper a security rationale as well as a developmental one: sovereign, regional, and distributed infrastructure is also a way of insulating civilian populations from other people's wars.

6.2 A development cooperation response

These challenges call for AI to be treated as a core development cooperation issue, not something left to a technology ministry. Five elements matter here:

1. **Finance the full stack, not stand-alone equipment.** Development banks and bilateral agencies should fund integrated packages of energy, connectivity, compute access, and skills together, rather than stand-alone AI projects. AI-ready infrastructure should qualify for concessional and climate-aligned finance, given its energy dimension. The upcoming replenishments of IDA and regional development funds are the practical vehicles for this.

2. **Compute as development infrastructure.** Support regional compute facilities under shared governance (a CERN model, adapted), national public-compute access programmes on the IndiaAI template, and negotiated access windows into existing infrastructure in partner countries. Concessional access to compute is becoming the twenty-first-century equivalent of concessional access to capital.

3. **Data and language as public goods.** Fund the creation of open, high-quality datasets in low-resource languages and development-critical domains such as health, agriculture and climate, building on existing digital public infrastructure efforts. Support national and regional data governance frameworks so developing-country data creates value at home instead of being extracted.

4. **Skills at scale, retention by design.** Move beyond short training workshops toward sustained investment: regional centres of excellence with guaranteed compute, research grants that can be held at home institutions, diaspora engagement schemes, South-South exchange. The aim is not only to train talent, but to make it viable for people to stay.

5. **Voice in governance.** Fund the participation of developing-country officials, researchers, and civil society in the Global Dialogue, in the Scientific Panel's consultations, and in technical standards bodies, including sustained representation in Geneva. Capacity-building for AI governance (model evaluation, incident response, regulatory design) should be a standing part of technical cooperation, drawing on the WSIS follow-up architecture that already connects digital development to this multilateral ecosystem.

Development cooperation of this kind is not an act of charity appended to the governance bargain described in Section 5; it is the delivery mechanism that makes the bargain credible in the first place. Access commitments negotiated in the Global Dialogue will be judged by whether they show up as funded, measurable programmes of exactly this sort.

7. Institutional Anchoring

The emerging UN machinery, the Global Dialogue on AI Governance and the Independent International Scientific Panel on AI, gives the binding core a legitimate, universal forum where it can be negotiated, and where the access bargain can be struck. The Scientific Panel is the natural custodian of the threshold-review function (Layer 1); the Dialogue is the natural venue for linking risk obligations to access commitments; and Geneva's standard-setting ecosystem (ITU, ISO/IEC liaison, the AI Safety Institute network) can host the interoperability layer.

8. Conclusion and Recommendations

This architecture holds under two conditions: the binding core has to stay genuinely narrow, and most of the world has to see that it serves shared safety and shared access rather than the interests of a few. That requires restraint from those with the greatest capabilities, and genuine, funded solidarity in return.

Recommendations:

1. Negotiate a binding international instrument narrowly scoped to extreme AI risks, using a periodically reviewed compute-and-capability threshold (10^{26}–10^{27} FLOPs plus demonstrated frontier capabilities) to define which systems it covers.

2. Vest threshold review in the Independent International Scientific Panel on AI, on a fixed cadence.

3. Build the interoperability layer now, without waiting for the binding instrument: common evaluation, incident-reporting, and documentation standards through the AI Safety Institute network, the OECD, and the ITU.

4. Link risk obligations and access commitments in a single negotiated bargain within the Global Dialogue on AI Governance, learning from the NPT's Article IV experience.

5. Fund open foundation models as global public goods through development banks and multilateral mechanisms, with Apertus as a proof of concept, and support regional compute pooling initiatives.

6. Mainstream AI into development cooperation: finance integrated packages (energy, connectivity, compute, skills) instead of stand-alone AI projects, make AI-ready infrastructure eligible for concessional finance, and fund open datasets in low-resource languages and development-critical domains.

7. Invest in talent retention and governance capacity in developing countries: regional centres of excellence with guaranteed compute, research funding that can be held at home institutions, and sustained support for developing-country participation in the Global Dialogue, the Scientific Panel, and technical standards bodies.

8. Launch a Montreux Document-style multistakeholder process on AI-enabled military operations, convening states, technology companies, and civil society in the tradition of Swiss-hosted humanitarian diplomacy, to clarify how international humanitarian law applies to mixed civilian-military digital infrastructure, and to develop shared standards for meaningful human judgment in AI-assisted targeting, complementing the CCW/LAWS track.

9. Preserve full national discretion over deployment, content, and industrial policy, and resist expanding the binding core beyond the identified extreme risks.

Sovereignty, safety and development need not be rivals. With careful design, and with restraint on all sides, each can strengthen the others.

References

1. António Guterres, UN Secretary-General, remarks to the United Nations Security Council on artificial intelligence and the maintenance of international peace and security, New York, 19 December 2024 (warning against the emergence of “fragmented AI spheres”; “A world of AI haves and have-nots would be a world of perpetual instability”).

2. T. Oweke, *The World Is Trying to Govern AI. The UN Wants In.*, Council on Foreign Relations, May 2026 — noting that more than eighty countries are advancing AI strategies or legislation based on differing regulatory philosophies, with attendant risks of fragmentation and transboundary harm.

3. United Nations High-level Advisory Body on Artificial Intelligence, *Governing AI for Humanity* — Final Report, September 2024, which found that 118 countries were party to none of the major international AI governance initiatives, and only seven developed countries were party to all.

4. Strategic Foresight Group (Sundeep Waslekar), *The Essential Convergence: Global Compact on Extreme AI Risks*, 2026. Launched in Geneva on 6 July 2026 by the Strategic Foresight Group, ICT4Peace Foundation, IFRC, and GCSP.

5. United Nations General Assembly, Resolution A/RES/79/325, *Establishment of the Independent International Scientific Panel on Artificial Intelligence and the Global Dialogue on Artificial Intelligence Governance*, August 2025.

6. United Nations General Assembly, Resolution A/RES/79/1, *The Pact for the Future* (including the Global Digital Compact), September 2024.

7. Global Dialogue on AI Governance, First Session, Geneva, 6–7 July 2026 — programme and documentation: https://www.un.org/global-dialogue-ai-governance/

8. European Union, *Regulation (EU) 2024/1689 (Artificial Intelligence Act)*, in particular the provisions on general-purpose AI models with systemic risk (Art. 51–55) and the 10^{25} FLOP presumption threshold.

9. United States, Executive Order 14110, *Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*, 30 October 2023 (rescinded January 2025), § 4.2 reporting threshold of 10^{26} FLOPs.

10. Basel Committee on Banking Supervision / Financial Stability Board, framework for Global Systemically Important Banks (G-SIBs): assessment methodology and higher loss-absorbency requirements, 2011, as updated.

11. Convention on International Civil Aviation (Chicago Convention), Chicago, 7 December 1944, establishing the International Civil Aviation Organization (ICAO) — the analogy most frequently invoked in AI governance debates.

12. R. Trager et al., *International Governance of Civilian AI: A Jurisdictional Certification Approach*, Oxford Martin School / Centre for the Governance of AI, 2023 — an "International AI Organization" modelled on ICAO, the IMO, and the FATF.

13. Treaty on the Non-Proliferation of Nuclear Weapons (NPT), 1968, in particular Article IV on the sharing of peaceful nuclear technology.

14. EPFL / ETH Zurich / Swiss National Supercomputing Centre (CSCS), *Apertus* — fully open Swiss large language model (open weights, data, and training methodology), 2025.

15. Government of India, *IndiaAI Mission* — national programme for public compute access and AI capacity, 2024.

16. International Telecommunication Union, *Facts and Figures* (latest edition) — global connectivity statistics, including the estimated one third of humanity remaining offline.

17. World Summit on the Information Society (WSIS), Geneva 2003 / Tunis 2005 outcome documents and the WSIS+20 review process, including the WSIS Forum 2026, Geneva, 6–10 July 2026.

18. ICT4Peace Foundation, *The Digital Transformation, Critical Minerals, and the Just Energy Transition*, January 2026.

19. Convention on Certain Conventional Weapons (CCW), Group of Governmental Experts on Lethal Autonomous Weapons Systems — rolling text (May 2025) and sessions of 2–6 March and 31 August–4 September 2026; Seventh CCW Review Conference, Geneva, 16–20 November 2026.

20. ICT4Peace Foundation (Anne-Marie Buzatu), *Bombing Clouds: Commercial Cloud Infrastructure, Military AI, and the Erosion of the Civilian-Military Distinction*, March 2026. https://ict4peace.org/wp-content/uploads/2026/03/Bombing-Clouds_ICT4Peace.pdf

21. *The Montreux Document on Private Military and Security Companies*, Swiss Federal Department of Foreign Affairs / ICRC, 2008 — precedent for a state-led, multistakeholder process clarifying the application of international humanitarian law to private actors.

ICT4Peace Foundation, Geneva