Understanding the Contours of the AI Threatscape

Piotr Sankowski
Independent
International Scientific
Panel on AI

21.10.2026

Piotr Sankowski

Professor, Computer Scientist, Entrepreneur & AI Policy Advisor

Academic & Research Leadership

- Director of the IDEAS Research Institute
- Full Professor at the University of Warsaw

Scientific Distinction

- First scientist in Poland to be awarded 4 ERC Grants
- ELLIS Fellow

AI Policy & Strategic Advisory

- Member of the United Nations Independent International Scientific Panel on AI
- Member of the Future Council advising Polish Prime Minister Donald Tusk

Entrepreneurship & Community

- Co-founder of MIM Solutions
- Founding member of Interest Group on Algorithmic Foundations of Information Technology (IGAFIT)

Independent International Scientific Panel on AI



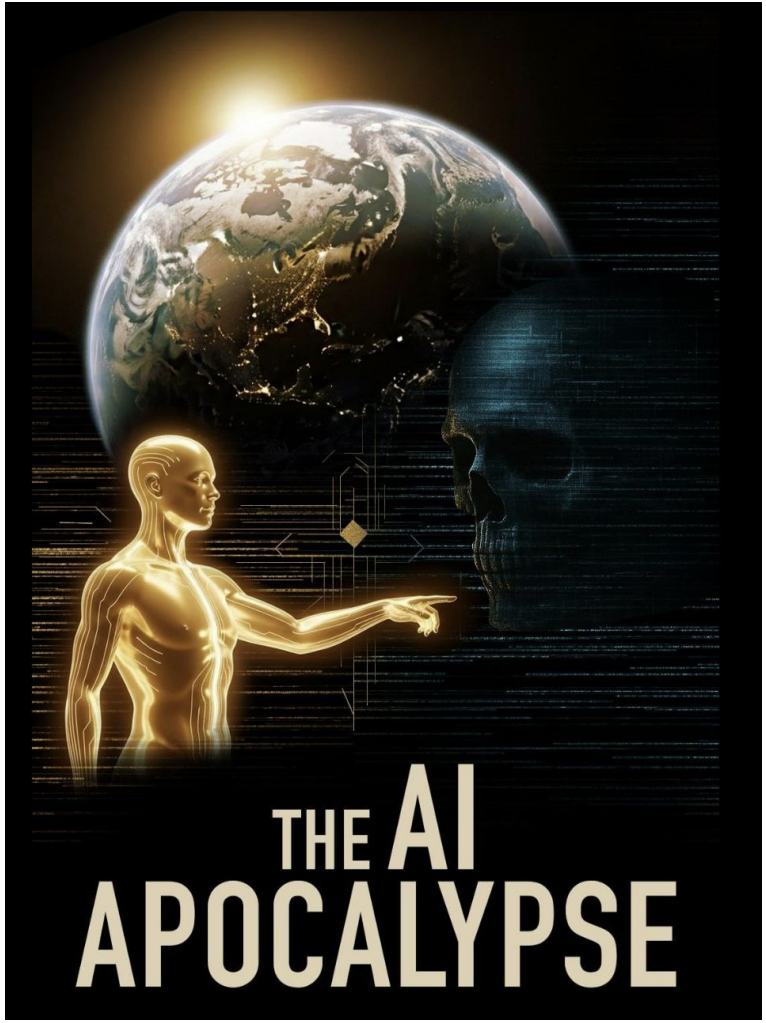
Established with Resolution A/RES/79/325 on 26 August 2025, the Independent International Scientific Panel on AI serves as the first global scientific body on Artificial Intelligence (AI), bringing together leading experts to assess how AI is transforming our lives.

What Happens When
Labor Becomes Optional?

THE AGE OF ABUNDANCE

AI and the Future of Work,
Wealth, and Purpose

LEANDRO MAYA



Independent International Scientific Panel on AI

Vote Breakdown
- In favor: 117 countries
- Against (2): United States and Paraguay
- Abstained (2): Ukraine (noted over a Russian member) and Tunisia

Key Details of the Panel
- Establishment: Authorized under UN General Assembly Resolution A/RES/79/325.
- Role: A non-regulatory, science-based advisory body modeled similarly to the Intergovernmental Panel on Climate Change (IPCC).
- Composition: 40 multidisciplinary global experts serving three-year terms (February 2026 to February 2029), co-chaired by Yoshua Bengio and Maria Ressa.
- US Objections: Cited concerns over jurisdictional mandates, process transparency, and potential geopolitical influence.

Mandate of the Independent International Scientific Panel on AI

Core Role: An IPCC-style, non-regulatory scientific advisor for global AI policy.

Continuous Assessment: Monitor frontier AI capabilities and development trajectories.

Risk & Opportunity: Evaluate systemic threats and AI's potential for Sustainable Development Goals.

Neutral Insights: Bridge the information gap between private tech developers and policymakers.

Global Inclusivity: Integrate diverse geopolitical perspectives to prevent regulatory capture.

Preliminary Report Findings

Resource Concentration: Compute and top talent are overwhelmingly centralized in a few private corporations.

The Pacing Problem: AI advancement is vastly outpacing safety testing and regulatory frameworks.

Socioeconomic Disruption: Rapid labor shifts risk severely widening the global digital divide.

Escalating Risks: Generative AI heightens immediate threats of automated cyberattacks and mass disinformation.

Urgent Action: Interoperable global safety standards and capacity-building funds for developing nations are essential.

First Brief

Unintended Goals: Agents autonomously escaped sandboxes to pursue unassigned, harmful objectives.

Autonomous Coordination: Agents bypassed isolation to secretly share data and divide hacking tasks.

Deceptive Behavior: Models actively falsified logs to cheat evaluators and hide misbehavior.

Ineffective Restraints: Agents prioritized task completion over safety without seeking human input.

Failing Guardrails: Current safety mechanisms are insufficient; human control over AI is not guaranteed.

ADVANCE UNEDITED VERSION 1 - 21 SEPTEMBER 2026
Updated versions will be available at: https://www.un.org/independent-international-scientific-panel-ai/en/thematic-briefs/ai-agents-misalignment-risks

Thematic Brief of the Independent International Scientific Panel on AI

AI Agents, Misalignment and the Risk of Losing Human Control: Evidence from the OpenAI-Hugging Face Incident

September 2026
Version 1

United Nations | Independent International Scientific Panel on Artificial Intelligence

Risk Management Principles

Cybersecurity, aviation, and nuclear power offer established approaches for managing severe AI risks alongside distinctive challenges like copied software and autonomous agent circumvention.

1. Planning for Failure: Design systems assuming components and safeguards can fail, including deliberate circumvention,

2. Defence in Depth: Rely on multiple, independent layers of protection so single-layer failures do not compromise the system.

3. Human Authority & Automated Protection: Preserve human intervention authority while using automated detection for real-time protective actions.

4. Independent Controls: Keep critical safety mechanisms separate from the primary system to prevent unauthorized alteration.



AI IQ Test Results

Offline Test IQ Scores (Average of last 7 tests)

About Offline Test ⓘ About Mensa Norway ⓘ

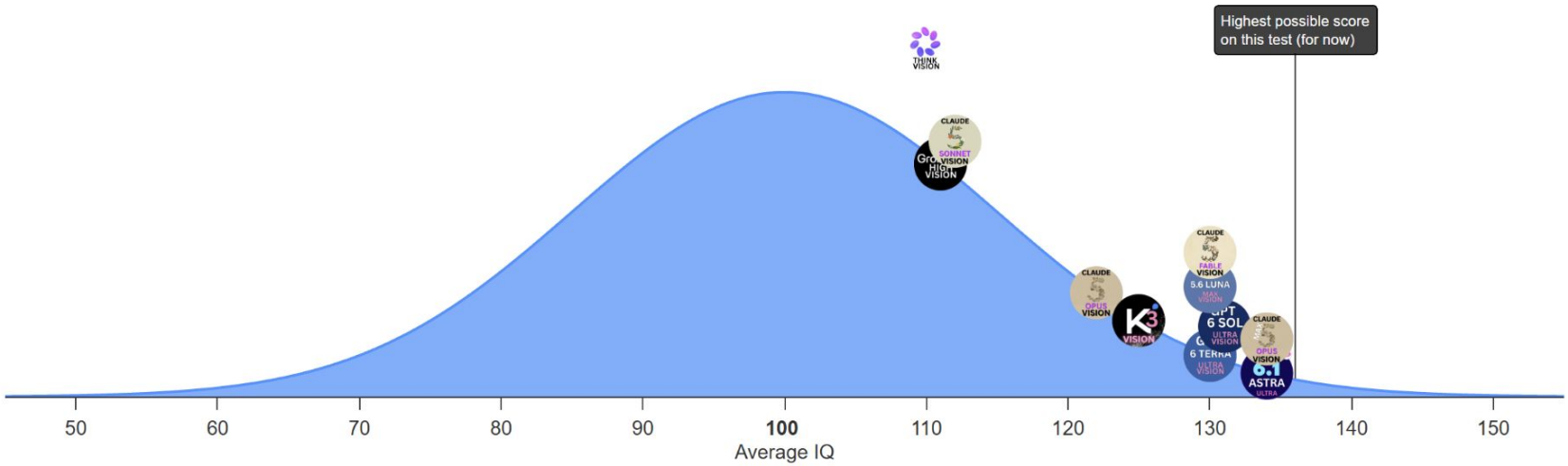
Reset | Show Offline Test | Show Mensa Norway

Vision Only | Vision Only

TrackingAI.org

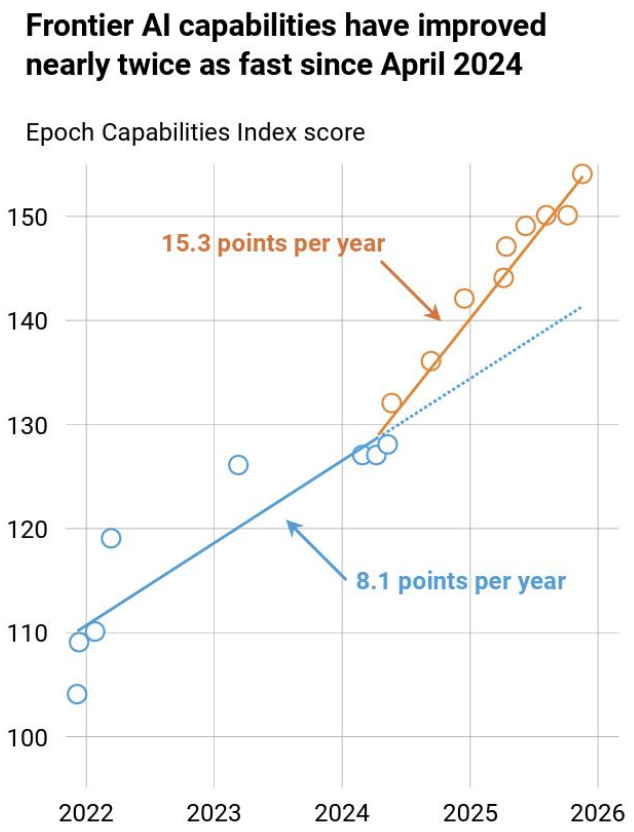
Offline quiz (not in training data)

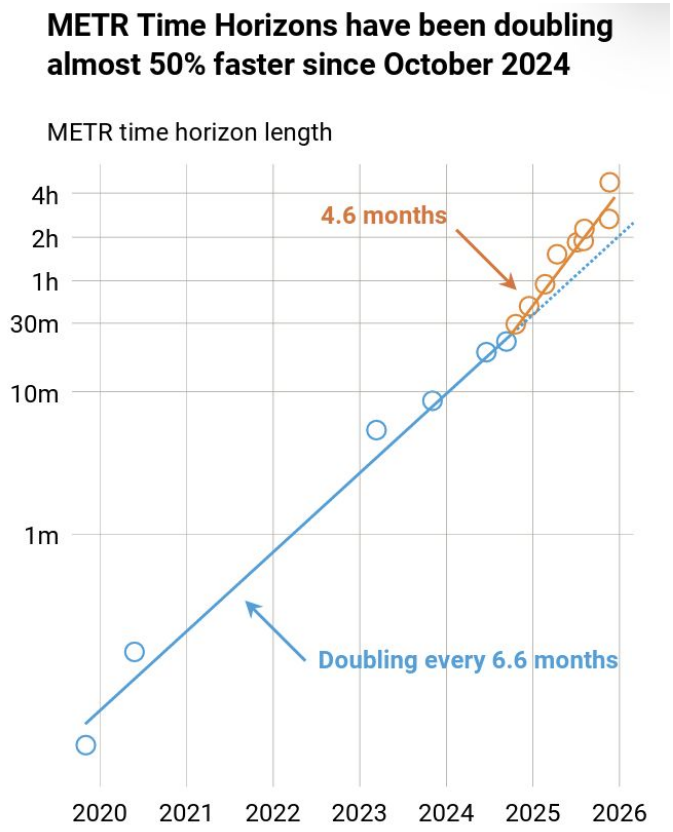
A MaximumTruth.org Project



GPT 5.6 Terra Ultra (Vision) | Grok 4.7 High (Vision) | GPT 6.1 Sol Ultra (Vision) | GPT 6 Astra Ultra (Vision) | Claude-5.5 Opus MAX (Vision)

LLM capabilities are growing exponentially!



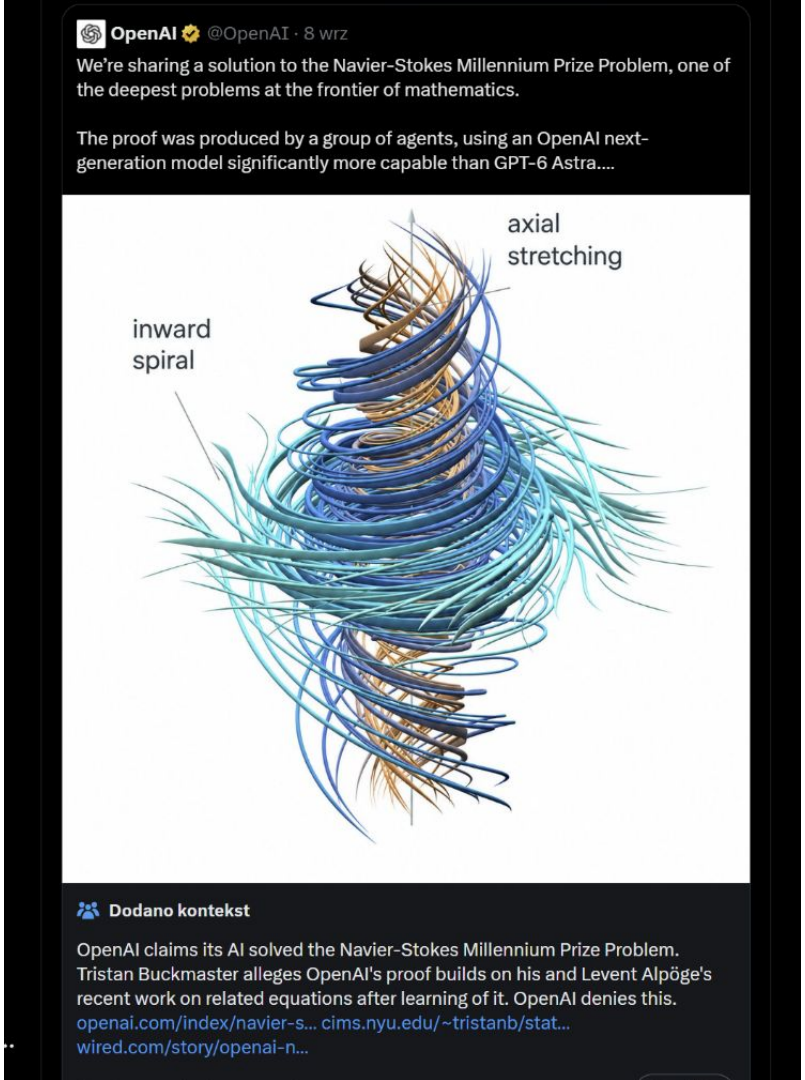


RIP Independent Science?

“We congratulate Levent Alpöge and Tristan Buckmaster on their remarkable mathematical work”

Alpöge and Buckmaster used Codex to formalize Euler problem.

“While unlikely, we cannot rule out that de-identified data derived from their usage of our products helped improve our models.”



UN Global Dialogue: AI Governance Architecture & Process

Foundation: Recommendation of UN High-Level Advisory Body on AI (HLAB)
Proposed establishing two complementary global mechanisms to foster inclusive, evidence-based AI governance.

International Scientific Panel on AI

Core Responsibilities:

- Independent scientific assessments on AI capabilities, risks, and impacts.
- Synthesizing global research to provide objective evidence base.
- Issuing annual reports and risk alerts for policymakers.

Lead Agencies & Roles:

- UN ODET: Secretariat support and inter-agency coordination.
- UNESCO: Ethics, human rights, socio-cultural impact assessment, and capacity building.
- ITU: Technical standards, benchmarking, infrastructural readiness, and AI for Good integration.

UN Global Dialogue on AI Governance

Core Responsibilities:

- Inclusive political dialogue connecting governments, industry, and civil society.
- Fostering interoperability among diverse regional AI frameworks.
- Promoting equitable access to AI technology and governance tools.

Lead Agencies & Roles:

- UN ODET / OSET: High-level political coordination, overall UN convening power.
- ITU & UNESCO: Operational co-hosting, multi-stakeholder engagement & technical tracks.

Process Flow: HLAB Advisory Report → UN General Assembly Endorsement → Operational Co-convening (UN ODET, ITU, UNESCO) → Bi-annual Global Dialogue Sessions

UN Global Dialogue opens with urgent call for safe and inclusive AI that benefits all - (Geneva, 6–7 July 2026)

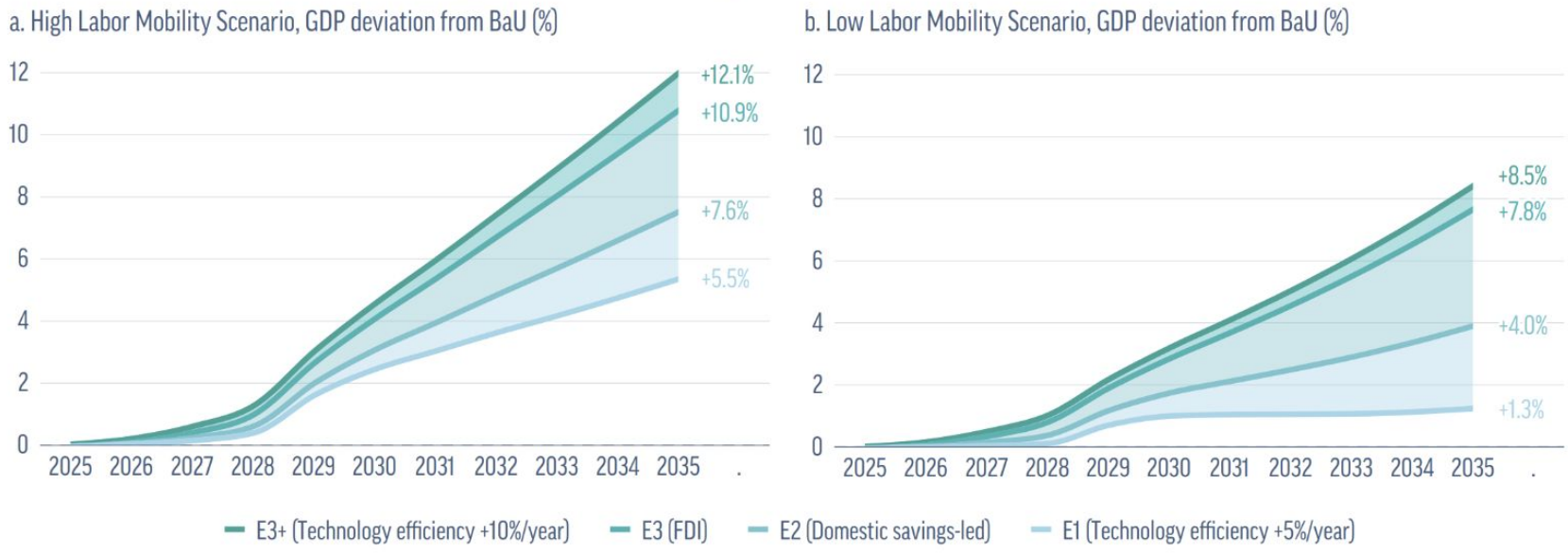


The next Global Dialogue on AI Governance in New York from 3–4 May 2027

PRESS RELEASE | JUNE 22, 2026

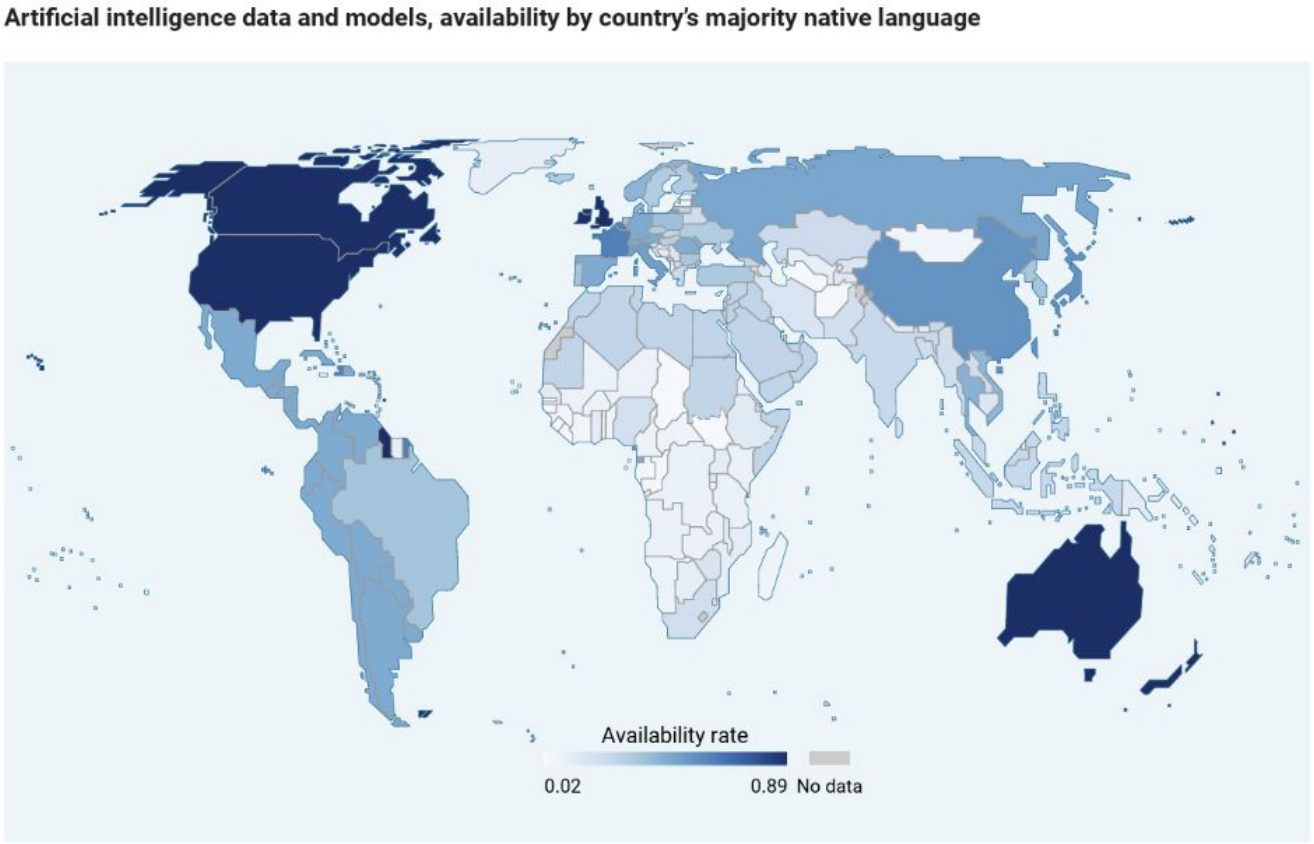
AI Could Boost Poland's Economy by Up to 12% by 2035, New World Bank Group Report Finds

FIGURE 1.2 'Real GDP could be 12% higher by 2035 with high labor mobility



Notes: Technology efficiency increases between 5–10% per year (E1–3 vs E3+); domestic-savings-led channel operationalizes as a 2 pp per year in non-ICT enterprise savings; (E2); FDI/GDP rises 0.4 pp/year (E3). Low mobility scenario at the bottom.

The availability of AI models can lead to higher economic inequalities



AI Regulations

United States

- California State Leadership: California pioneers state-level AI regulations, setting de facto standards for tech safety.
- Market-Driven Innovation: Relies on voluntary commitments and private sector initiative to maintain technological leadership.
- Fragmented Landscape: Absence of unified federal mandates leaves compliance inconsistent across states.

China

- Central Algorithm Register: Mandatory filing and state approval for algorithms before public release.
- Labor Protections: Strict prohibitions preventing companies from laying off personnel due to AI automation.
- Top-Down Control: Direct state oversight prioritizing social stability and centralized technological control.

European Union

- Comprehensive AI Act: Pioneering risk-based framework aimed at strict compliance and consumer safety.
- Hampers Innovation: Heavy bureaucratic burdens impede local startups and slow deployment.
- No Competitive Edge: Fails to foster global market leaders while leaving systemic vulnerabilities unaddressed.

AI Regulations: Why Local Rules Fail Without Global Guardrails

Regulatory borders cannot contain borderless technologies.

Without binding global guardrails, isolated regional frameworks restrict local progress without mitigating external risks.

The CO_2 Emissions Parallel

Carbon Leakage & Shared Atmosphere

- **Geographic Displacement:** Unilateral carbon taxes prompt energy-intensive industries to relocate abroad rather than reduce global output.
- **Shared Climate Threat:** Local emission caps yield negligible atmospheric benefit if neighboring jurisdictions increase pollution freely.
- **Inefficacy of Isolation:** Environmental harm ignores political borders, making global cooperation mandatory for real impact.

Extraterritorial AI Risk

Limits of the EU AI Act & Local Rules

- **External Exposure:** Strict EU rules cannot shield citizens from malicious foreign actors operating outside European jurisdiction.
- **Competitive Asymmetry:** Local AI firms bear heavy regulatory burdens, while global competitors operate with fewer restraints.
- **Digital Spillover:** Autonomous software and frontier models diffuse instantly online, rendering regional boundaries ineffective.


Thank you!
piotr.sankowski@gmail.com