BBIMUN'26

FUTURE OF AI SUMMIT

BACKGROUND GUIDE

Balancing Innovation, Regulation and Global Equity in the Governance of Artificial Intelligence

TABLE OF CONTENTS

1. Letter from the Executive Board

Greetings Delegates,

Welcome to the Blue Bells International Model United Nations and to the Future of AI Summit. This committee addresses and deals with upcoming developments and issues that are related to artificial intelligence.

We, being the Executive Board members, would like to show gratitude to the organizers for letting us be a part of this wonderful opportunity. We hope that by now your research is well on its way and you have formulated an idea about what, how, why, and when you want to discuss something. These questions form the very basis of the flow of debate and argumentation in the committee. This background guide will give you an overview of the topic at hand and the work of the Committee. It contains some basic elements on the topic that will guide your research. However, such mentions do not limit the scope of discussion in the committee at all.

We expect from all delegates an active participation in the proceedings of this committee to have a fruitful discussion on a pertinent global problem. For that purpose, extensive and thorough research is expected of you over and beyond this study guide. Think of this study guide as merely an initiation to your research, defining the broad aspects. A section with the questions to ponder has been added at the end of this document and can be utilized to that regard.

Also, an important point here is that while criticism is encouraged, we expect constructive criticism in the committee. This will help you approach a problem differently and understand all perspectives. What needs to be noted here is that opinions are only different, but never wrong. An opinion is a product of multiple factors that an individual is exposed to during the course of life and hence needs to be deconstructed and not discarded.

While forming your arguments, take logical premises and not ludicrous ones as an argument is only as strong as the premise it is based on. Try to communicate your premises, followed by your arguments and then a conclusion keeping in mind the time limit. This will help you convey your message effectively. Many delegates make a mistake of quoting only articles or resolutions without explaining their relevance in the agenda. Be rational like why this resolution is important or what a particular article of a convention implies to structure your research better.

Your research should always answer "What" ,"Why" ,"How" and "When" , A common mistake Delegates make is simply quoting facts and figures(which can also be found on the internet), and making a logical premise out of it. Try to link these facts and figures and then direct your research. We wish you all the very best!

Vanshaj Arora
(Chairperson)

2. The Nature of Proof/Evidence in Committee

Evidence or proof is acceptable from the following sources

→ **News Sources**:

State operated News Agencies: These reports can be used in the support of or against the State that owns the News Agency. These reports, if credible or substantial enough, can be used in support of or against any Country as such but in that situation, they can be denied by any other country in the council. Some examples are -

1. IRNA (Iran) http://www.irna.ir/ENIndex.htm,

2. BBC (United Kingdom) http://www.bbc.co.uk/

3. Xinhua News Agency and CCTV (P.R. Of China) http://cctvnews.cntv.cn/

→ **Government Reports**:

These reports can be used in a similar way as the State Operated News Agencies reports and can, in all circumstances, be denied by another country. However, a nuance is that a report that is being denied by a certain country can still be accepted by the Executive Board as credible information.

Examples are Government Websites like:

1. State Department of the United States of America: http://www.state.gov/index.htm ,

2. Ministry of Defense of the Russian Federation: http://www.eng.mil.ru/en/index.htm ,

3. Permanent Representatives to the United Nations Reports:

http://www.un.org/en/members/ (Click on any country to get the website of the Office of its Permanent Representative.)

4. Multilateral Organizations like the NATO

(http://www.nato.int/cps/en/natolive/index.htm) ASEAN

(http://www.aseansec.org/), OPEC (http://www.opec.org/opec_web/en/), etc.

→ **UN Reports**:

All UN Reports are considered as credible information or evidence for the Executive Board of the General Assembly.

1. UN Bodies: Like the SC (http://www.un.org/Docs/sc/), GA (http://www.un.org/en/ga/), HRC (http://www.ohchr.org/EN/HRBodies/HRC/Pages/HRCIndex.aspx) etc.

2. UN Affiliated bodies like the International Atomic Energy Agency (http://www.iaea.org/), World Bank (http://www.worldbank.org/), International Monetary Fund (http://www.imf.org/external/index.htm), International Committee of the Red Cross (http://www.icrc.org/eng/index.jsp, etc.

3. Treaty Based Bodies like the Antarctic Treaty System (http://www.ats.aq/e/ats.htm), the International Criminal Court (http://www.icc-cpi.int/Menus/ICC) .

Under no circumstances will sources like Wikipedia (http://www.wikipedia.org/), Amnesty International (http://www.amnesty.org/) or newspapers like the Guardian (http://www.guardian.co.uk/), Times of India (http://timesofindia.indiatimes.com/) etc. be accepted as credible.

3. Executive Summary

Artificial intelligence has moved, within roughly three years, from a specialist research field to a general-purpose technology treated by heads of state, central banks, and militaries as a first-order strategic asset. The release of ChatGPT in November 2022 triggered a global "AI boom" in investment, adoption, and regulatory activity that shows no sign of slowing: Stanford HAI's 2026 AI Index records global corporate AI investment of $581.7 billion in 2025 (more than double the prior year), generative AI adoption reaching an estimated 53 percent of the world's population within three years of ChatGPT's release, and AI incident reports rising to 362 documented cases. At the same time, the gap between the most and least capable AI ecosystems has widened: high-income countries account for the overwhelming majority of notable model releases and AI investment, while many low-income countries hold a negligible share of global data-centre compute capacity.

The governance response has been fast but fragmented. No binding multilateral treaty on AI exists. Instead, delegates face a patchwork of binding regional law (chiefly the EU's Artificial Intelligence Act, Regulation (EU) 2024/1689), binding national law of varying stringency (China's Interim Measures for the Management of Generative AI Services; a growing body of U.S. state legislation), non-binding multilateral declarations (Bletchley, Seoul, the Paris and New Delhi statements), voluntary corporate safety frameworks (Anthropic's Responsible Scaling Policy, OpenAI's Preparedness Framework, and the G7 Hiroshima Code of Conduct signed by sixteen-plus firms), and a set of nascent United Nations mechanisms (the Independent International Scientific Panel on AI and the Global Dialogue on AI Governance, both established by General Assembly resolution in August 2025). These layers frequently pull in different directions: the United States under the second Trump administration has pursued deregulation and pre-emption of state AI laws in the name of maintaining U.S. "AI dominance," while the European Union has continued to implement binding, risk-tiered obligations even as it has granted industry a phased delay on the toughest requirements.

This Background Guide equips delegates to negotiate a text responsive to three simultaneous demands that define the agenda: sustaining the pace of innovation that underwrites economic competitiveness and scientific progress; building regulatory and safety guardrails proportionate to genuine risks, from algorithmic discrimination to loss-of-control scenarios in frontier models; and ensuring that the economic and scientific benefits of AI, and the infrastructure that produces them, are not

monopolised by a handful of firms and states. The Committee should expect no ready consensus. The Bletchley (2023), Seoul (2024), Paris (2025), and New Delhi (2026) summits each produced weaker or narrower agreement than the last on binding commitments, even as participation broadened. Delegates representing the United States and China will resist any text that constrains national compute or export-control policy; delegates representing the European Union and civil-society-aligned researchers will push for enforceable safety and transparency standards; and delegates representing the Global South will centre digital public infrastructure, compute access, and equitable data governance. A workable outcome document will likely combine narrow, technically achievable commitments (interoperable incident reporting, a shared vocabulary for risk tiers, expanded compute-sharing mechanisms) with a roadmap for revisiting harder questions such as binding safety testing or a global compute registry.

.

4. Committee Mandate

4.1 Scope of Influence: Advisory v. Binding

The Future of AI Summit is convened as a multi-stakeholder deliberative body, not a treaty-making conference. Its outputs are recommendatory rather than binding on states, corporations, or international organisations; the Committee cannot amend the EU AI Act, alter U.S. executive orders, or compel Chinese regulators to change the Interim Measures. This mirrors the actual architecture of global AI governance to date: every multilateral AI instrument produced since 2023 , the Bletchley Declaration, the Seoul Declaration and Frontier AI Safety Commitments, the Hiroshima Process International Code of Conduct, the Paris "Statement on Inclusive and Sustainable Artificial Intelligence," and the New Delhi Declaration on AI Impact , has been explicitly non-binding, resting on political commitment rather than legal obligation. The Committee's comparative advantage lies precisely here: because its resolutions carry no direct legal force, delegates can propose more ambitious or more granular language than national negotiators typically risk in binding fora, provided the substance is realistic enough that governments, firms, and civil-society organisations named in any resolution could plausibly act on it.

Delegates should nonetheless treat the distinction between binding and non-binding instruments as analytically central throughout debate. A resolution that "calls upon" the EU AI Office to share enforcement data operates on fundamentally different footing than a clause that "urges" Anthropic or OpenAI to publish red-teaming results under their existing Responsible Scaling Policy or Preparedness Framework commitments, since the latter are voluntary corporate undertakings the companies can revise unilaterally (as OpenAI did to its Preparedness Framework in April 2025, narrowing the risks it evaluates before release). Conflating hard law, soft law, corporate self-regulation, and aspirational rhetoric in a single operative clause is a common and avoidable drafting error.

4.2 Dynamics Between State and Non-State Actors

Unlike a conventional UN General Assembly committee, this Summit's stakeholder matrix places sitting heads of state and government alongside corporate chief executives, practising researchers (, and civil-society advocates. This composition reflects reality: frontier AI capability, compute infrastructure, and safety expertise are concentrated overwhelmingly in a small number of private firms rather than governments, a point the UN Secretary-General's High-Level Advisory Body on AI made

centrally in its September 2024 report "Governing AI for Humanity." Delegates should expect state and corporate interests to align on some questions and diverge sharply on others.

4.3 Voting and Consensus Mechanisms

Consistent with the practice of every AI summit convened to date, this Committee should default to consensus-based adoption of its final declaration, with a roll-call vote available where consensus is not achievable on discrete operative clauses. Delegates should study the precedent set at the Paris AI Action Summit (February 2025), where the United States and United Kingdom declined to sign the closing "Statement on Inclusive and Sustainable Artificial Intelligence" even though 58-60 other countries did, and at the New Delhi Declaration on AI Impact (February 2026), which secured endorsement from 88-89 countries and organisations, including, notably, both the United States and China, but which achieved this breadth partly by keeping commitments voluntary and framed around shared principles ("Sutras") rather than specific obligations. A Committee resolution that only the Global South signs, or only the transatlantic bloc signs, would represent a governance failure by the standard the agenda itself sets: global equity requires buy-in that spans the compute-rich and compute-poor alike.

5. Agenda Overview

The agenda, "Balancing innovation, regulation and global equity in the governance of Artificial Intelligence", deliberately yokes together three objectives that are not always mutually reinforcing. Innovation policy rewards speed, capital concentration, and permissive experimentation; regulation of the kind embodied in the EU AI Act rewards predictability, auditability, and precaution; and global equity requires redistributing compute, data, talent, and decision-making power away from the small set of incumbents, overwhelmingly U.S. and Chinese firms and, to a lesser extent, states, that currently hold them. Delegates should organise their preparation around four linked questions that recur throughout every session of this Background Guide:

- Where should the line sit between voluntary corporate self-governance (Responsible Scaling Policy, Preparedness Framework, the Hiroshima Code of Conduct) and binding law (the EU AI Act, China's Interim Measures, prospective U.S. federal or state statutes)?
- How should states balance frontier-AI safety commitments against the competitive pressure to deploy quickly, especially amid the US-China compute race and the collapse of the "safety summit" consensus after Bletchley?
- What legal and economic mechanisms, licensing, data dividends, compute-sharing, digital public infrastructure, can meaningfully close the gap between AI "haves" and "have-nots" without freezing the current leaders' advantage in place?
- What institutional home, if any, should global AI governance eventually have: a UN-anchored body (building on the Independent International Scientific Panel on AI and the Global Dialogue on AI Governance), a G7/G20-style club, an IAEA-style verification body ("CERN for AI"), or continued reliance on parallel national and regional regimes?

Delegates negotiating on behalf of states should be prepared to reconcile their government's stated domestic AI policy with the position they advance in Committee; delegates representing corporate or civil-society figures should be prepared to justify positions by reference to those figures' publicly documented statements, safety frameworks, or advocacy record rather than invented positions.

6. Historical Background

6.1 From Turing to Transformers to Frontier Models

Modern AI governance debates rest on six decades of technical history that delegates should be able to summarise concisely rather than recite in full. Alan Turing's 1950 paper "Computing Machinery and Intelligence" and the 1956 Dartmouth workshop (where the term "artificial intelligence" was coined) mark the field's founding. Symbolic, rules-based AI dominated research through the 1980s, followed by two "AI winters" of reduced funding driven by unmet expectations. The 2012 "deep learning" breakthrough (the AlexNet result in the ImageNet competition) reignited investment by demonstrating that large neural networks trained on large datasets with sufficient compute could outperform hand-engineered systems. The 2017 publication of the "Transformer" architecture by Google researchers (Vaswani et al., "Attention Is All You Need") supplied the technical foundation for the large language models (LLMs) that now dominate public and policy attention. OpenAI's GPT series (2018-2023), Google DeepMind's work following its 2023 merger, and Anthropic's Claude models (Anthropic was founded in 2021 by former OpenAI researchers including Dario Amodei) all build on transformer-based architectures scaled with unprecedented compute and data.

6.2 Historical Timeline

Three factors specifically explain why AI governance became a first-order political issue only from late 2022 onward, rather than earlier in this six-decade history. First, ChatGPT's public release on 30 November 2022 was the first frontier AI product built for, and marketed directly to, consumers at massive scale, making capabilities and risks legible to legislators and the general public in a way earlier research systems were not. Second, the same period saw the emergence of credible warnings from senior AI researchers , including a March 2023 open letter (signed by figures including Elon Musk) calling for a pause on giant AI experiments, and subsequent public statements by 2024 Nobel physics laureate Geoffrey Hinton, who left Google in 2023 citing AI safety concerns , that generated political urgency disproportionate to any single technical benchmark. Third, the compute, capital, and talent required to train frontier models concentrated rapidly in a small number of firms (OpenAI, Google DeepMind, Anthropic, Meta, xAI, and, in China, firms such as DeepSeek, Alibaba, and Zhipu), converting AI policy into industrial and great-power competition policy almost immediately.

6.3 Major Turning Points

Date	Event	Significance
Nov 2022	Public release of ChatGPT	First mass-market generative AI product; catalyses global regulatory attention
Mar 2023	Italian Garante temporarily bans ChatGPT	First Western state enforcement action against a generative AI service
May 2023	G7 launches the Hiroshima AI Process	First G7-level coordinated framework on advanced AI systems
Jul-Aug 2023	China's Interim Measures for Generative AI Services take effect	First binding national law targeting generative AI specifically
Oct 2023	Biden signs Executive Order 14110 on AI; China unveils Global AI Governance Initiative	Competing US and Chinese framings of international AI governance
Nov 2023	UK hosts the Bletchley Park AI Safety Summit; Bletchley Declaration signed by 28 countries plus the EU	First head-of-government-level multilateral statement on frontier AI risk
Mar 2024	UN General Assembly adopts Resolution A/RES/78/265 by consensus	First UN resolution dedicated to AI, sponsored by the US and 123 co-sponsors
May 2024	Seoul AI Summit; Seoul Declaration and Frontier AI Safety Commitments	16+ companies commit to publish safety frameworks; national AI Safety Institute network launched
Sep 2024	UN High-Level Advisory Body publishes "Governing AI for Humanity"	Proposes UN-anchored scientific panel and global dialogue on AI
Jan 2025	Trump rescinds EO 14110; issues EO 14179 "Removing Barriers to American Leadership in AI"	US pivots from risk-mitigation to deregulation-first AI policy

Date	Event	Significance
Feb 2025	Paris AI Action Summit; US and UK decline to sign closing declaration	First major rupture in the Bletchley-Seoul-Paris safety-summit consensus
Aug 2025	UNGA Resolution A/RES/79/325 establishes the Scientific Panel on AI and Global Dialogue on AI Governance	First standing UN institutional machinery for AI
Sep 2025	Bartz v. Anthropic $1.5bn settlement granted preliminary approval	Largest reported copyright recovery in history; first major AI-training copyright resolution
Nov 2025	UK High Court rules for Stability AI in Getty Images v. Stability AI	First major UK judgment on AI training and copyright/trade mark law
Feb 2026	India AI Impact Summit adopts the New Delhi Declaration on AI Impact (89 endorsers, incl. US and China)	Broadest endorsement of any AI declaration to date; first Global South-hosted AI summit

7. Current Situation

7.1 Latest Developments

As of mid-2026, several trends dominate the landscape. Frontier model capability has continued to advance rapidly on standard benchmarks , Stanford HAI's 2026 AI Index records performance on the SWE-bench Verified coding benchmark rising from roughly 60 percent to near 100 percent of the human baseline within a single year, and reports that the top four frontier models (from Anthropic, xAI, Google, and OpenAI) were separated by fewer than 25 Elo points on independent model-ranking leaderboards as of March 2026 , while transparency has fallen: the Foundation Model Transparency Index tracked by the AI Index dropped from 58 to 40 points as leading labs, including OpenAI, Anthropic, and Google, stopped disclosing training compute, dataset composition, and parameter counts for their newest systems. Regulatory momentum in the EU has slowed relative to its original timetable: the "Digital Omnibus" reform, provisionally agreed in May 2026 and in force from late July 2026, deferred the AI Act's toughest high-risk obligations (Annex III systems) from August 2026 to December 2027, and Annex I embedded systems to August 2028, while leaving the transparency and general-purpose-AI oversight provisions on their original August 2026 schedule. In the United States, the Trump administration has continued to prioritise deregulation and federal pre-emption, issuing a further executive order in December 2025 directing an "AI Litigation Task Force" to challenge state AI laws deemed to conflict with national policy.

7.2 Compute Bottlenecks and the Semiconductor Supply Chain

Advanced-AI compute remains choke-pointed at a small number of firms, above all NVIDIA (whose data-centre GPUs , the Hopper-generation H100/H200 and Blackwell-generation B100/B200/GB200 lines , dominate frontier model training) and the Taiwan Semiconductor Manufacturing Company (TSMC), which fabricates nearly all leading-edge AI chips for NVIDIA and its rivals, including AMD. Export controls on advanced chips to China, first imposed by the U.S. Commerce Department's Bureau of Industry and Security in October 2022 and repeatedly revised since, have become one of the most consequential and volatile instruments of AI governance. The Trump administration reversed elements of the prior "AI diffusion" framework and, in a series of steps beginning December 2025, authorised licensed sales of NVIDIA's H200 chip (though not the more advanced Blackwell line) to a limited set of vetted Chinese buyers subject to a 25 percent

revenue-sharing arrangement with the U.S. Treasury and case-by-case Commerce Department review. As of mid-2026, actual shipments remained minimal: Chinese regulators had not cleared bulk imports, reportedly to protect the market position of domestic alternatives such as Huawei's Ascend line, even as Chinese firms placed large speculative orders. NVIDIA disclosed that despite licences covering roughly ten major Chinese customers, it had recognised no China data-centre revenue from H200 sales as of its early-2026 earnings calls. Independent analysts estimate the United States retains a 20-to-50-times advantage in AI compute production over China even accounting for licensed H200 flows, though China's domestic chip industry (Huawei, SMIC, Cambricon) continues to narrow the gap and to benefit politically from the disruption U.S. export-control reversals have caused among U.S. allies and firms.

7.3 Regional Dynamics

The US-China relationship remains the central axis of AI geopolitics, oscillating between confrontation (export-control tightening, entity-list additions, antitrust scrutiny of NVIDIA in China) and negotiated de-escalation (the H200 licensing deal, agreed around the same period as a Trump-Xi summit in 2026). China has pursued a dual strategy of pressing for a China-centred multilateral AI body , Xi Jinping's October 2023 Global AI Governance Initiative and his November 2025 APEC proposal for a "World AI Cooperation Organization" headquartered in Shanghai, explicitly framed as an alternative to a U.S.-led order , while simultaneously accelerating domestic chip self-sufficiency. The European Union has positioned itself as the "regulatory frontrunner" through the AI Act, but faces a widening competitiveness gap: EU officials and outside commentators (including a Snowflake policy analysis cited widely in 2026) note sustained industry pressure for further delay, and von der Leyen's own €200 billion "InvestAI" pledge (announced at the Paris summit alongside a €20 billion AI "gigafactories" fund) signals recognition that regulation alone will not secure European competitiveness. The Global South, largely absent from the Bletchley and Seoul summits' most consequential negotiations, gained unprecedented convening power at the February 2026 India AI Impact Summit , the first global AI summit hosted in the Global South , which produced the broadest-endorsed AI declaration to date (89 countries and organisations) built around seven thematic "Chakras," including a voluntary, non-binding "Charter for the Democratic Diffusion of AI" aimed at broadening access to foundational AI resources.

8. International Legal Framework

8.1 Binding National and Regional Law

8.1.1 The European Union AI Act (Regulation (EU) 2024/1689)

Official title and year: Regulation (EU) 2024/1689 of the European Parliament and of the Council, laying down harmonised rules on artificial intelligence, entered into force 1 August 2024. The Act is the first comprehensive, horizontal AI statute in the world and applies extraterritorially to any AI system placed on the EU market or whose output is used in the EU, regardless of where it is developed.

Relevant structure: the Act sorts AI systems into four risk tiers.

1. **Unacceptable risk (Article 5):** outright prohibited practices, including social scoring by public authorities, manipulative or exploitative techniques, untargeted scraping of facial images, and (with narrow law-enforcement exceptions) real-time remote biometric identification in public spaces; in force since 2 February 2025, with two additional prohibitions (including AI-generated child sexual abuse material) added from 2 December 2026 under the 2026 Digital Omnibus reform.

2. **High risk (Article 6, Annexes I and III):** systems used in domains such as biometrics, critical infrastructure, education, employment, essential services, law enforcement, migration, and justice; subject to risk management, data governance, technical documentation, logging, human oversight, conformity assessment, and CE marking obligations.

3. **Limited/transparency risk (Article 50):** disclosure duties such as informing users they are interacting with an AI system and labelling AI-generated or manipulated media; in force since 2 August 2026 with no delay.

4. **Minimal risk:** no specific obligations beyond a general AI-literacy duty (Article 4). General-purpose AI (GPAI) models are governed separately under Chapter V, with transparency and, for models presenting "systemic risk," additional obligations that took effect 2 August 2025.

Why it matters and current status: penalties reach €35 million or 7 percent of global annual turnover for Article 5 violations, higher than the equivalent GDPR ceiling. Under the "Digital Omnibus"

(Regulation (EU) 2026/1744, provisionally agreed 7 May 2026 and in force from 27 July 2026), the Article 6/Annex III high-risk obligations originally due 2 August 2026 were deferred to 2 December 2027, and Annex I embedded high-risk systems to 2 August 2028, following sustained industry lobbying that the original timetable was unworkable; the transparency duties and the AI Office's enforcement powers over GPAI providers were not delayed and remain in force from 2 August 2026. Delegates should note this delay is a live, contested fact: it is neither a full retreat from binding regulation nor evidence the Act's core architecture (the four-tier system, conformity assessment, and GPAI oversight) has changed.

8.1.2 China's Interim Measures for the Management of Generative AI Services

Official title and year: Interim Measures for the Management of Generative Artificial Intelligence Services, jointly issued 10 July 2023 by the Cyberspace Administration of China (CAC) and six other ministries, effective 15 August 2023 , China's first binding regulation targeting generative AI specifically, and among the earliest such national regulations anywhere.

Relevant articles: Article 4 requires that generative AI content uphold "core socialist values" and prohibits content that incites subversion, undermines national unity, or promotes terrorism, extremism, or discrimination; Article 17 requires security assessment and algorithm filing (with the Cyberspace Administration) for services with "public opinion properties or the capacity for social mobilisation"; Article 22 defines "generative AI technology"; Article 24 sets the effective date. The Measures were formulated under, and sit alongside, China's Cybersecurity Law, Data Security Law, and Personal Information Protection Law. Complementary rules include the 2022 Deep Synthesis Provisions (governing "deepfake" content) and, from 1 September 2025, mandatory content-labelling rules ("Labeling Rules") requiring implicit (metadata) and, where perceptible to users, explicit labelling of AI-generated text, audio, image, video, and virtual-scene content.

Why it matters: the Measures combine a "inclusive and prudent," innovation-friendly posture toward developers with strict content controls and mandatory security review before public deployment , an approach explicitly distinct from both the EU's harmonised risk-tiering and the U.S.'s sectoral, deregulation-leaning approach. Enforcement has been active: by within weeks of the Measures taking effect, eleven major Chinese generative AI services (including Baidu's Ernie Bot and Alibaba's Tongyi Qianwen) had completed the required algorithm filing to operate publicly, and China's State Council has continued to develop a broader national "Artificial Intelligence Law," although as of the freeze

date no comprehensive AI statute beyond the sector-specific Measures and related provisions had been finalised.

8.1.3 United States: Executive Orders and the Absence of Federal Legislation

The United States has no single comprehensive federal AI statute; governance instead rests on a shifting patchwork of executive orders, sectoral guidance, and state law. President Biden's Executive Order 14110, "Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence" (30 October 2023), directed federal agencies to develop AI safety-testing guidance, created reporting requirements for large training runs, and established a U.S. AI Safety Institute within the Department of Commerce. On his first day back in office (20 January 2025), President Trump revoked EO 14110 via Executive Order 14148 ("Initial Rescissions of Harmful Executive Orders and Actions"). On 23 January 2025, Trump issued Executive Order 14179, "Removing Barriers to American Leadership in Artificial Intelligence," declaring it U.S. policy to "sustain and enhance America's global AI dominance" and directing preparation of an AI Action Plan, released 23 July 2025. On 11 December 2025, Trump issued a further order, "Ensuring a National Policy Framework for Artificial Intelligence," directing the Attorney General to establish an AI Litigation Task Force to challenge state AI laws deemed to conflict with federal policy on grounds including unconstitutional burdens on interstate commerce and federal preemption , a direct response to the growing body of state-level AI legislation (all 50 states introduced AI-related bills in 2025, and 38 had adopted transparency or safety measures by year's end, according to reporting cited in industry coverage of Anthropic's public advocacy on the issue).

Why it matters: the reversal from EO 14110 to EO 14179 illustrates how quickly binding-adjacent federal AI policy can change with no legislative involvement, and how directly it can conflict with the EU's more durable, legislatively enacted framework. It also explains why voluntary corporate frameworks and state law have become comparatively more important checks on frontier AI development in the U.S. than federal rulemaking.

8.2 Multilateral Soft Law

The OECD AI Principles (adopted 2019, updated 2024) were the first intergovernmental AI policy framework and remain the normative foundation cited by nearly every subsequent instrument, including the G7 Hiroshima Process. The Hiroshima AI Process, launched by the G7 in May 2023, produced two documents on 30 October 2023: the "International Guiding Principles for

Organizations Developing Advanced AI Systems" and the voluntary "International Code of Conduct for Organizations Developing Advanced AI Systems," the latter of which more than sixteen companies (including Amazon, Anthropic, Google, Meta, Microsoft, OpenAI, and, from Seoul onward, NVIDIA and xAI) had committed to implement by the 2024 Seoul Summit. The OECD subsequently piloted a voluntary reporting framework to monitor Code of Conduct implementation (2024). The Bletchley Declaration (1 November 2023, 28 countries plus the EU) was the first head-of-government-level acknowledgment that frontier AI could cause "serious, even catastrophic harm." The Seoul Declaration and accompanying Frontier AI Safety Commitments (21-22 May 2024) built on Bletchley by securing commitments from named frontier developers to publish safety frameworks and by launching an international network of AI Safety Institutes. The Paris "Statement on Inclusive and Sustainable Artificial Intelligence for People and the Planet" (11 February 2025) was endorsed by 58-60 countries but not by the United States or United Kingdom, marking the first major split in the safety-summit consensus. The New Delhi Declaration on AI Impact (20-21 February 2026), adopted at the India AI Impact Summit, secured the broadest endorsement of any AI declaration to date (89 countries and international organisations, including both the U.S. and China) by organising commitments around seven voluntary, non-binding thematic pillars ("Chakras") and a voluntary "Charter for the Democratic Diffusion of AI."

8.3 Corporate Self-Regulation and Frontier AI Safety Commitments

Anthropic's Responsible Scaling Policy (RSP), first published September 2023 and now in its fourth substantive revision (Version 3.1, effective 2 April 2026), is a voluntary public commitment not to train or deploy models above defined "AI Safety Level" (ASL) capability thresholds without matching deployment and security safeguards. Version 3.0 (24 February 2026) introduced published "Frontier Safety Roadmaps" and annual "Risk Reports" subject to external review. OpenAI's Preparedness Framework, first published December 2023 and substantially revised in April 2025, evaluates models against "high" and "critical" capability thresholds across categories including cybersecurity, biological/chemical risk, and model autonomy; the 2025 revision drew criticism from safety researchers, including a former OpenAI employee, for narrowing the framework to no longer specifically evaluate persuasion/manipulation risk and for weakening commitments to test fine-tuned model variants. Because both frameworks are voluntary and unilaterally revisable, they function as soft governance: they create reputational and (for RSP, contractually referenced) commitments but no

external enforcement mechanism, and their substantive content can and does change as competitive pressure evolves.

8.4 Intellectual Property and Copyright Law Applicability

No jurisdiction has enacted AI-specific copyright legislation resolving whether training generative models on copyrighted works constitutes infringement; the question is being litigated under existing fair-use (U.S.), text-and-data-mining exception (EU/UK), and general copyright doctrine. The World Intellectual Property Organization (WIPO) has convened member-state conversations on AI and IP since 2019 but has not produced a binding instrument. The WTO has not adopted AI-specific trade rules, though AI training data flows and chip export controls increasingly intersect with existing trade-law obligations, a tension several WTO members have raised informally but not litigated to a ruling as of the freeze date. Section 7 below details the landmark litigation shaping this area.

9. Landmark Cases

9.1 The New York Times Company v. Microsoft Corporation and OpenAI, Inc. et al.

- **Citation:** The New York Times Company v. Microsoft Corp. et al., No. 1:23-cv-11195 (S.D.N.Y., filed 27 December 2023), Judge Sidney H. Stein, consolidated with related publisher actions (MDL 1:25-md-03143).

- **Facts:** the Times alleges OpenAI and Microsoft used millions of Times articles without authorisation to train GPT models powering ChatGPT and Copilot, and that the models can reproduce Times content near-verbatim, undermining its subscription business; claims include direct and contributory copyright infringement and DMCA violations.

- **Judgment/current status:** on 26 March 2025, Judge Stein largely denied OpenAI and Microsoft's motion to dismiss, allowing the bulk of the case to proceed toward a possible jury trial; as of mid-2026 the case remains in contested discovery, with a January 2026 order (affirmed by Judge Stein) compelling OpenAI to produce a sample of 20 million ChatGPT conversation logs, and a July 2026 sanctions motion by the Times and other publishers alleging OpenAI concealed its ability to search training data and output logs for infringing content , allegations OpenAI disputes. No trial date has been set.

- **Legal principle:** whether large-scale ingestion of copyrighted text for LLM training is "transformative" fair use notwithstanding potential market substitution effects and near-verbatim regurgitation.

- **Importance for the agenda:** the case is widely regarded as the most consequential U.S. test of AI-training fair use and bears directly on whether journalism, and creative industries generally, can extract compensation from frontier AI developers , a central global-equity and IP question for the Committee.

9.2 Bartz el al. v. Anthropic PBC

- **Citation:** Bartz et al. v. Anthropic PBC, No. 3:24-cv-05417 (N.D. Cal.), Judge William Alsup, filed August 2024 by authors Andrea Bartz, Charles Graeber, and Kirk Wallace Johnson on behalf of a certified class.

- **Facts:** plaintiffs alleged Anthropic downloaded millions of books from pirate "shadow libraries" (affecting an estimated 500,000 titles) to train its Claude models.

- **Judgment/current status:** in June 2025, Judge Alsup issued a split ruling , holding that Anthropic's use of lawfully acquired books to train Claude was protected, transformative fair use, but that Anthropic's retention of more than seven million pirated books in a "central library" independent of any training use violated the authors' rights, exposing Anthropic to statutory damages for the piracy itself rather than for the training use. Facing a December 2025 damages trial with potentially ruinous exposure, Anthropic agreed in September 2025 to a $1.5 billion class settlement (approximately $3,000 per affected title), which Judge Alsup preliminarily approved as "fair, reasonable, and adequate" later that month; the settlement does not grant Anthropic a forward-looking licence and requires destruction of the pirated files.

- **Legal principle:** this is the first substantive U.S. judicial ruling to separate two distinct questions squarely , whether training itself is fair use (yes, on these facts) and whether the manner of acquiring training data is independently unlawful (yes, where the works were pirated) , and produced what plaintiffs' counsel called the largest publicly reported copyright recovery in history.

- **Importance for the agenda:** the ruling gives AI developers a partial fair-use victory on training while signalling that provenance of training data carries independent legal risk, a distinction directly relevant to any Committee recommendation on data-sourcing standards.

9.3 Getty Images (US) Inc. & Ors. v. Stability AI Ltd.

- **Citation:** Getty Images (US) Inc & Ors v. Stability AI Limited [2025] EWHC 2863 (Ch), High Court of England and Wales, judgment handed down 4 November 2025.

- **Facts:** Getty alleged that Stability AI's Stable Diffusion image-generation model was trained (largely outside the UK) on millions of Getty's copyrighted images without licence, and

separately alleged trade mark infringement based on the model occasionally reproducing a distorted version of the Getty watermark.

- **Judgment/current status:** the Court dismissed Getty's primary copyright claims, holding that because the training and development of Stable Diffusion occurred outside the UK, and the trained model weights are not themselves a form in which the copyright works are stored or reproduced, there was no "importation" of an infringing copy into the UK when the pre-trained model was later made available there; the Court found only an "extremely limited," now-historical trade mark infringement relating to watermark reproduction in early model outputs.

- **Legal principle:** models trained abroad do not, by that fact alone, infringe UK copyright when later deployed in the UK, unless the model itself stores or reproduces the protected works , a significant limitation on the territorial reach of copyright claims against AI developers who train outside the claimant's jurisdiction.

- **Importance for the agenda:** the judgment is the first substantive UK ruling on AI training and copyright/trade mark law and illustrates how territorial and technical questions (where training occurs; what a model's weights legally "are") can determine liability regardless of the underlying policy merits , reinforcing why several governments, including the UK, have separately opened consultations on whether copyright law needs AI-specific reform.

9.4 Garante per la Protezione dei Dati Personali (Italian Data Protection Authority) v. OpenAI

- **Citation:** proceedings before the Garante per la Protezione dei Dati Personali (Italy), commencing with an emergency order of 30-31 March 2023 and concluding with a sanctions decision issued 20 December 2024.

- **Facts:** on 31 March 2023, the Garante ordered OpenAI to immediately stop processing the personal data of individuals in Italy, provisionally banning ChatGPT in Italy , the first such action by a Western regulator against a generative AI service , citing the absence of a legal basis for the mass collection of personal data used to train ChatGPT, a March 2023 data breach OpenAI had not reported, and inadequate age-verification safeguards for users under 13.

OpenAI restored service in Italy by end of April 2023 after implementing remedial measures, including an EU-user opt-out mechanism and clearer privacy disclosures.

- **Judgment/current status:** on 20 December 2024, the Garante concluded its underlying investigation and fined OpenAI €15 million, finding that OpenAI processed personal data to train ChatGPT without an appropriate legal basis, breached transparency obligations, failed to notify the March 2023 breach, and lacked adequate age-verification mechanisms; it also ordered a six-month public information campaign. OpenAI called the fine "disproportionate" (noting it exceeded roughly twenty times OpenAI's reported Italian revenue in the relevant period) and stated it would appeal.

- **Legal principle:** the case establishes that GDPR's legal-basis and transparency requirements apply squarely to the collection and use of personal data for LLM training and to the operation of a public-facing chatbot, and that "legitimate interest" claims by AI developers are not a self-executing defence.

- **Importance for the agenda:** it is the first generative-AI enforcement action brought under the GDPR and a template other EU data-protection authorities and the AI Act's emerging enforcement bodies are likely to draw on.

10. Multilateral and UN Actions

10.1 Major UN General Assembly Resolutions

Resolution A/RES/78/265, "Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development," was adopted without a vote by the General Assembly on 21 March 2024, sponsored by the United States and 123 co-sponsoring states, including China and others with whom the U.S. negotiated the text over several months. It is a framework resolution: it does not create binding obligations but calls on member states to promote AI systems that are safe, secure, and trustworthy, to close AI-related digital divides, and to ensure developing countries can access AI's benefits (disease detection, flood prediction, agricultural support, workforce training). Resolution A/RES/79/1 ("Pact for the Future," adopted September 2024), through its annexed Global Digital Compact, mandated the creation of two follow-on mechanisms, formally established by Resolution A/RES/79/325 on 26 August 2025: the Independent International Scientific Panel on Artificial Intelligence, a 40-member multidisciplinary body providing annual, policy-relevant but non-prescriptive scientific assessments of AI capabilities and risks (its first cohort was recommended by the Secretary-General for General Assembly appointment in 2026), and the Global Dialogue on AI Governance, an annual inclusive forum for states and stakeholders, with its first session convened in Geneva in July 2026 and a second scheduled for New York in 2027.

10.2 Agency Mandates: ITU and UNESCO

The International Telecommunication Union (ITU), led by Secretary-General Doreen Bogdan-Martin, coordinates the annual "AI for Good" summit and technical standardisation work on AI, and participates in the UN's broader AI governance machinery; ITU does not, however, hold binding regulatory authority over AI content or safety, only over the telecommunications standards that increasingly intersect with AI infrastructure. UNESCO adopted the Recommendation on the Ethics of Artificial Intelligence in November 2021 , the first global normative instrument on AI ethics, unanimously adopted by all 194 member states , which sets non-binding principles including proportionality, human oversight, and environmental sustainability, and which UNESCO has since operationalised through a voluntary "Readiness Assessment Methodology" used by dozens of countries to benchmark their AI governance capacity.

10.3 High-Level Advisory Body Reports

The UN Secretary-General's High-Level Advisory Body on Artificial Intelligence (HLAB-AI), convened 26 October 2023 with 39 members from 33 countries serving in a personal capacity, published an interim report in December 2023 and its final report, "Governing AI for Humanity," on 19 September 2024. The final report's central finding is a set of governance gaps , representational (most of the world has no voice in frontier AI governance), jurisdictional (no institution has a mandate spanning AI's full risk spectrum), and coordination-related (existing initiatives are fragmented) , addressed through seven recommendations, including the scientific panel and global dialogue mechanisms subsequently adopted in Resolution A/RES/79/325, a proposed AI standards exchange, a capacity-development network, a global fund to close compute and data gaps, and a small UN policy dialogue and secretariat. Delegates should treat "Governing AI for Humanity" as the single most authoritative UN-system analytical document underpinning this agenda and consult it directly (un.org/en/ai-advisory-body) when drafting resolution language.

11. Stakeholder Analysis

11.1 Sovereign States and National Security Interests

The United States (represented in the matrix by President Donald Trump and Vice President J.D. Vance) has, since January 2025, pursued a deregulation-first, "AI dominance" posture: revoking Biden-era risk-mitigation policy, releasing a July 2025 AI Action Plan oriented around infrastructure build-out and export competitiveness, and moving in December 2025 to preempt state-level AI regulation. Vance's remarks at the Paris summit ("pro-growth AI policies" over "excessive regulation") and the U.S. and UK's refusal to sign the Paris declaration crystallised this posture internationally. China (Xi Jinping) has combined domestic content-and-security control (the Interim Measures, mandatory content labelling) with an assertive international governance push , the October 2023 Global AI Governance Initiative and the November 2025 proposal at APEC for a Shanghai-headquartered "World AI Cooperation Organization" , explicitly framed as an alternative to U.S.-centred governance, while racing to reduce dependence on U.S. chip exports. India (Narendra Modi) has positioned itself as convenor of the Global South through the IndiaAI Mission and the February 2026 India AI Impact Summit, emphasising digital public infrastructure, compute democratisation, and "AI for all," while also courting frontier-lab and chip-industry investment (including a semiconductor manufacturing facility Modi inaugurated in Uttar Pradesh during the Summit). The European Union (Ursula von der Leyen) and France (Emmanuel Macron) have pursued a dual-track strategy of binding regulation (the AI Act, now partially delayed) paired with large public-private investment pledges (Macron's €109 billion February 2025 package and von der Leyen's €200 billion "InvestAI" initiative, including €20 billion for AI "gigafactories"), reflecting an explicit judgment that Europe cannot out-regulate its way to AI competitiveness. The United Kingdom (Keir Starmer) has, notwithstanding having hosted the original Bletchley Summit, opted for a "pro-innovation," light-touch regulatory posture that avoided both a comprehensive AI act and the Paris declaration, instead extending existing sectoral regulators' remit to cover AI. France's junior digital minister Anne Le Hénanff represents the operational, implementation-level face of French and EU digital policy, including AI Act enforcement coordination and France's national AI strategy execution.

11.2 The Compute Providers

NVIDIA (Jensen Huang) supplies the overwhelming majority of the specialised GPUs used to train frontier models and is the direct object of U.S.-China export-control policy; Huang has lobbied actively (including meetings with Trump ahead of the December 2025 H200 policy reversal and a 2026 visit to China) for continued market access, arguing restricted access accelerates Chinese self-sufficiency rather than containing it, while U.S. export-control advocates and members of Congress argue the opposite. NVIDIA's market capitalisation (roughly $5 trillion by mid-2026) and its centrality to virtually every frontier lab's training infrastructure make it a de facto governance chokepoint independent of any formal regulatory authority. AMD (Lisa Su) is NVIDIA's principal competitor in AI accelerators and has similarly sought export licences (its MI308 chip was cleared for China alongside NVIDIA's H20/H200) while investing heavily in expanding capacity to meet global demand. Apple (John Ternus) and IBM (Arvind Krishna) sit at the compute and enterprise-infrastructure periphery of this debate , Apple through on-device/edge AI and its comparatively cautious public AI rollout, IBM through enterprise AI governance tooling and its long-standing advocacy for "precision regulation" targeted at high-risk use cases rather than broad model-level rules.

11.3 Frontier Model Developers

OpenAI (Sam Altman, with Mira Murati a notable departure who left in 2024 to found Thinking Machines Lab) completed a contentious corporate restructuring in October-November 2025, converting its capped-profit subsidiary into a Delaware public benefit corporation while keeping the nonprofit OpenAI Foundation in a controlling, roughly 26-percent-equity position , a structure critics (including former employees who filed briefs in Elon Musk's related litigation against OpenAI) argue dilutes the nonprofit's ability to enforce its founding safety mission. OpenAI has simultaneously narrowed its voluntary Preparedness Framework (April 2025) and pursued the Stargate infrastructure joint venture (with SoftBank and Oracle, announced January 2025, targeting up to $500 billion in U.S. AI infrastructure investment). Anthropic (Dario Amodei; Jack Clark, co-founder and policy lead) has staked out a comparatively pro-regulation position among frontier developers: Amodei has publicly stated he is "deeply uncomfortable" with AI's trajectory being decided by "a few companies, by a few people" (CBS 60 Minutes, November 2025) and, in 2025, called a proposed decade-long moratorium on state AI regulation "too blunt an instrument" in a New York Times op-ed, instead urging a federal transparency standard developed jointly by Congress and the White House. Anthropic's Responsible Scaling Policy (now Version 3.1) is the most detailed and most frequently revised of the major voluntary safety frameworks. Google DeepMind (Demis Hassabis, co-recipient of the 2024 Nobel

Prize in Chemistry for protein-structure prediction work) operates within Alphabet's (Sundar Pichai) broader AI strategy and has generally supported measured, internationally coordinated safety regulation while continuing rapid frontier deployment (Gemini). xAI (Elon Musk) has combined public warnings about AI existential risk , Musk co-signed the March 2023 open letter calling for a pause on giant AI experiments and called AI "one of the biggest threats to humanity" at the Bletchley Summit , with a company strategy of rapid, comparatively less safety-encumbered model releases (Grok) and Musk's parallel litigation against OpenAI alleging its restructuring betrays its founding nonprofit mission, a stance critics note sits uneasily beside his own commercial AI ambitions. Microsoft (Satya Nadella; Brad Smith, Vice Chair and President; Mustafa Suleyman, CEO of Microsoft AI following the 2024 acquisition of much of Inflection AI's talent) is both OpenAI's largest investor/infrastructure partner and a co-defendant alongside OpenAI in the New York Times litigation, giving it a distinctive dual exposure as both frontier-AI financier and regulatory target. Safe Superintelligence Inc. (Ilya Sutskever, OpenAI co-founder and former chief scientist, who departed OpenAI in 2024 following the November 2023 board crisis) and Cohere (Aidan Gomez, a co-author of the original 2017 Transformer paper) represent smaller, differently positioned frontier and enterprise-AI developers with generally research-safety-oriented public postures.

11.4 Open-Source Champions

Meta (Mark Zuckerberg; Joel Kaplan, Chief Global Affairs Officer) has marketed its Llama model family as open-weight, arguing broad access accelerates safety research, competition, and Global South adoption; this stance faces internal strain following Meta's 2025 pivot toward a more closed, "Superintelligence Lab" model-development structure under new leadership (including hires from Scale AI and OpenAI) and the high-profile November 2025 departure of long-time chief AI scientist and open-source advocate Yann LeCun, who left to found a new "world models" research venture (AMI Labs) after voicing sustained skepticism that scaling large language models will reach advanced AI and frustration with Meta's LLM-focused strategic direction. Hugging Face (Thomas Wolf, co-founder and Chief Science Officer) operates the world's largest open-model and open-dataset hosting platform and has consistently argued, including in EU AI Act consultations, that open-source/open-weight development lowers barriers for researchers, smaller states, and the Global South, and improves auditability relative to closed frontier systems, while acknowledging this same openness raises proliferation risks for misuse. Stability AI (Emad Mostaque, founder, who stepped down as CEO in March 2024 amid company turmoil) remains a reference point for open-weight image

generation and was the losing defendant on the primary claim, but not the narrow trade-mark claim, in Getty Images v. Stability AI.

11.5 Civil Society and Safety Advocates

Max Tegmark (Future of Life Institute) has campaigned for binding international limits on frontier AI development and helped organise the March 2023 pause letter; FLI also maintains a widely cited EU AI Act compliance-checking tool. Tristan Harris (Centre for Humane Technology) focuses on AI's societal and attentional harms (addiction, misinformation, erosion of shared reality) and has testified before national legislatures on algorithmic amplification. Geoffrey Hinton, the 2024 Nobel physics laureate and one of deep learning's founding figures, left Google in 2023 specifically to be able to speak freely about AI risk and has since become one of the most prominent establishment voices warning of both near-term harms and longer-term loss-of-control risk; Yoshua Bengio (a fellow "godfather of AI") has played a parallel role chairing the International AI Safety Report process. Andrew Ng (DeepLearning.AI) and, more skeptically, Yann LeCun represent the strongest establishment counter-voices, arguing that existential-risk framing is overstated, distracts from addressable present-day harms (bias, labour disruption, concentration), and risks entrenching incumbents through regulatory capture. Amandeep Singh Gill (UN Secretary-General's Envoy on Technology) has led the UN Global Digital Compact and Scientific Panel/Global Dialogue process from within the Secretariat. Kai-Fu Lee, a prominent China-based AI investor and former Google China head, brings a distinct US-China comparative-capability perspective, having written extensively on the two countries' divergent AI development models. Nandan Nilekani (Infosys chairman and architect of India's Aadhaar digital-identity system) is the leading global advocate for digital public infrastructure (DPI) as a governance and equity tool, arguing India's DPI model , population-scale digital identity, payments (UPI), and data-sharing consent layers , offers a template other countries, particularly in the Global South, can adapt for AI-era governance, and was a featured voice at the India AI Impact Summit's sessions on governing AI within digital public infrastructure.

12. Key Issues

12.1 Existential Risk v. Immediate Harms

A persistent fault line separates researchers and advocates who prioritise long-horizon, catastrophic or existential risk (loss of control over highly capable autonomous systems, AI-enabled bioweapons or cyberattacks) , a camp associated with Hinton, Bengio, Tegmark, and, institutionally, Anthropic's ASL framework and OpenAI's "critical capability" thresholds , from those who argue this framing is speculative, empirically ungrounded, and diverts attention and regulatory capacity from documented present-day harms: algorithmic bias and discrimination in hiring, credit, and criminal justice; large-scale misinformation and synthetic media; labour displacement; and environmental costs of data-centre energy and water use. Delegates should note this is not simply an academic dispute: it maps onto concrete policy choices, such as whether safety regulation should gate frontier training runs above a compute threshold (implying existential-risk framing) versus regulating specific deployed use-cases regardless of the underlying model's scale (the EU AI Act's dominant logic, and IBM's long-standing "precision regulation" advocacy).

12.2 The Open-Source Debate: Proliferation Risk v. Democratisation

Open-weight models (Meta's Llama family, several Chinese labs' releases including DeepSeek's, and Mistral AI in France) lower the cost of AI adoption for researchers, startups, and states without frontier-lab budgets, and allow independent safety auditing impossible with closed weights. Critics , including some frontier-lab safety teams and national-security-oriented policymakers , counter that open-weighting a sufficiently capable model is effectively irreversible proliferation: once weights are public, no subsequent safety patch, usage restriction, or export control can be applied retroactively, a concern sharpened by China's early-2025 release of the highly capable, low-cost DeepSeek-R1 model, which demonstrated that state-of-the-art capability no longer requires U.S.-scale compute budgets and briefly triggered a sharp reassessment of U.S. AI-stock valuations. The Hiroshima Code of Conduct and most Seoul Frontier AI Safety Commitments apply irrespective of release strategy, but neither instrument resolves whether, or at what capability threshold, open-weighting should itself require the kind of pre-release safety testing currently expected only of closed frontier deployments.

12.3 Intellectual Property and Data Scraping

U.S. courts have so far treated AI training on lawfully acquired copyrighted material as likely transformative fair use (Bartz v. Anthropic) while leaving the underlying acquisition of that material (piracy, unlicensed scraping) independently actionable, and UK courts have limited copyright claims' territorial reach where training occurs abroad (Getty v. Stability AI). No jurisdiction has adopted a comprehensive AI-specific compensation or licensing regime for training data, though the EU AI Act's GPAI transparency obligations require providers to publish a "sufficiently detailed summary" of training content and to respect the EU's existing text-and-data-mining opt-out regime under the 2019 Copyright Directive. The core policy tension is unresolved: rights-holders (publishers, authors, visual artists, represented in ongoing litigation by organisations such as the Authors Guild and the News/Media Alliance) argue AI developers extract enormous commercial value from copyrighted human creative and journalistic labour without consent or compensation, while developers argue that broad liability for training would functionally halt frontier AI development outside jurisdictions willing to tolerate it (implicitly advantaging China, where copyright enforcement against AI training is comparatively undeveloped).

12.4 Global Equity and Digital Public Infrastructure

Stanford HAI's 2026 AI Index finds high-income countries account for roughly 87-91 percent of notable AI model production and startup funding, while low-income countries collectively hold roughly 0.1 percent of global data-centre compute capacity , a gap the UN High-Level Advisory Body's "Governing AI for Humanity" identified as a central governance failure. Two broad remedial strategies dominate current debate. The Digital Public Infrastructure (DPI) approach, championed by India and Nandan Nilekani and formally referenced in the New Delhi Declaration's voluntary "Charter for the Democratic Diffusion of AI," argues that population-scale, interoperable, publicly governed digital identity, payments, and data-consent layers (modelled on India's Aadhaar and UPI systems, which Nilekani states are already being adapted in roughly twenty countries with an ambition to reach fifty) can let lower-income states capture AI's productivity benefits without needing to build sovereign frontier-model capacity themselves. The compute-and-capacity-sharing approach, reflected in proposals discussed at the UN Global Dialogue on AI Governance and the New Delhi Summit's "AI Commons" and "shared compute infrastructure" working-group deliverables, argues that direct redistribution or pooling of training compute, cloud credits, and technical capacity-building is necessary because DPI alone does not address the underlying dependency on a handful of chip and cloud providers concentrated in the United States and China.

12.5 Compute Governance and Export Controls

U.S. export-control policy toward China has become one of the most volatile and consequential instruments in global AI governance, cycling between the Biden administration's tiered "AI diffusion" framework, its rescission, and the Trump administration's licensed, revenue-shared H200 sales , a policy China's own regulators have partly blocked to protect domestic chip-makers. Beyond the US-China axis, "compute governance" as a broader concept , tracking, licensing, or capping large training runs above a defined compute threshold, an approach with structural similarities to nuclear-material safeguards , remains almost entirely unimplemented outside the EU AI Act's narrower GPAI "systemic risk" compute threshold (10^25 FLOP) and voluntary frontier-lab self-reporting. No international body currently verifies compute usage or training-run characteristics across jurisdictions.

13. Statistics

The following figures are drawn from the most recent editions of the sources indicated and should be cited with year and source in any resolution or position paper; several (particularly investment and adoption figures) are estimates subject to methodological revision, and delegates should flag genuine uncertainty rather than treat any single figure as definitive.

Metric	Figure	Source / Year
Global corporate AI investment, 2025	$581.7 billion (total corporate investment incl. M&A/minority stakes); $344.7bn private investment alone	Stanford HAI AI Index 2026
Private AI investment growth, 2025 vs. 2024	+127.5%	Stanford HAI AI Index 2026
Generative AI population-level adoption	~53% globally within three years of ChatGPT's release (faster than the PC or internet); U.S. ranks 24th at 28.3% adoption	Stanford HAI AI Index 2026
Foundation Model Transparency Index	Fell from 58 to 40 (out of 100) year-on-year; 80 of 95 notable 2025 models shipped without training code	Stanford HAI AI Index 2026
Documented AI incidents, 2025	362 reported incidents	Stanford HAI AI Index 2026
Notable AI models released, 2025, by country	United States 59-60; China 35	Stanford HAI AI Index 2026 / Epoch AI
U.S. consumer surplus from generative AI, annualised, early 2026	~$172 billion (up from ~$112bn a year earlier)	Stanford HAI AI Index 2026

Metric	Figure	Source / Year
Software developer employment, ages 22-25, change since 2024	-20%	Stanford HAI AI Index 2026
Global employment exposed to AI	~40% (60% in advanced economies; 40% emerging markets; 26% low-income countries)	IMF, January 2024
Projected net global jobs created by AI/technology, 2025-2030	170 million created; 92 million displaced; net +78 million (~14% of current employment)	WEF Future of Jobs Report 2025 (Jan. 2025)
Bartz v. Anthropic settlement value	$1.5 billion (~$3,000 per affected title; ~500,000 titles)	N.D. Cal. Case No. 3:24-cv-05417, Sept. 2025
Italian Garante fine on OpenAI	€15 million	Garante decision, 20 Dec. 2024
EU AI Act maximum penalty (Art. 5 prohibited practices)	€35 million or 7% of global annual turnover, whichever is higher	Regulation (EU) 2024/1689, Art. 99
France AI investment pledge (Macron, Feb. 2025)	€109 billion (~$112bn)	Reuters/TechCrunch, Feb. 2025
EU "InvestAI" pledge (von der Leyen, Feb. 2025)	€200 billion, incl. €20bn for AI "gigafactories"	European Commission, Feb. 2025
New Delhi Declaration on AI Impact endorsements	89 countries and international organisations (as of 22 Feb. 2026)	Government of India, Feb. 2026

Metric	Figure	Source / Year
Estimated US compute production advantage over China (2026, even with licensed H200 flows)	~21-49x, depending on metric	Analyst estimates cited in techjournal.org, 2026
High-income countries' share of notable AI models / startup funding	~87% / ~91%	Stanford HAI AI Index 2026 (developing-nations comparative analysis)
Low-income countries' share of global data-centre compute capacity	~0.1%	World Bank / Stanford HAI 2026 comparative analysis

Insufficient reliable evidence exists to state a single, uncontested figure for the compute cost of training any specific current frontier model (OpenAI, Anthropic, and Google have all stopped disclosing compute and dataset details for their newest systems, per the Foundation Model Transparency Index decline noted above); delegates should treat any such figure cited in debate as an estimate and request its source and methodology before relying on it in resolution language.

14. Challenges and Gaps

14.1 Legal: Jurisdictional Evasion

The EU AI Act's extraterritorial reach, the U.S.'s sectoral patchwork, and China's content-and-security-first model create three incompatible compliance regimes that a single global firm must navigate simultaneously, and the Getty v. Stability AI judgment demonstrates how straightforward it can be to structure training activity (conducting it outside a claimant jurisdiction) to limit legal exposure. States and firms alike can and do engage in regulatory arbitrage: training runs, data acquisition, and even deployment infrastructure can be routed through permissive jurisdictions, and chip-export-control evasion (the "Blackwell loophole" the U.S. Commerce Department moved to close in mid-2026, involving Chinese-headquartered firms' subsidiaries outside China) shows enforcement consistently lags circumvention.

14.2 Political: Geopolitical Weaponisation of AI

AI governance has become inseparable from great-power competition: U.S. export controls are simultaneously industrial policy, non-proliferation policy, and diplomatic leverage (the H200 licensing decision was reportedly discussed in the context of a broader Trump-Xi trade negotiation); China's Global AI Governance Initiative and World AI Cooperation Organisation proposal are explicitly framed as counters to a U.S.-led order; and the collapse of the Bletchley-Seoul-Paris safety-summit consensus tracks the broader deterioration of transatlantic policy alignment after January 2025. This weaponisation complicates the Committee's task directly: any resolution language on compute-sharing, export controls, or a global verification body will be read by major-power delegates as either advantaging or disadvantaging their state's strategic position, regardless of its stated humanitarian or scientific rationale.

14.3 Economic: Monopolisation by Big Tech Giants

Frontier AI development requires capital and compute at a scale (Stargate's proposed $500 billion; Meta's reported plans to exceed $100 billion in single-year AI spending) that only a handful of firms and states can plausibly mobilise, and the Stanford HAI concentration statistics in Section 11 show this is already producing a durable structural gap rather than a transitional one. Vertical integration , NVIDIA's dominance of both chips and, increasingly, cloud/AI-service offerings; Microsoft's

position as both OpenAI's primary infrastructure provider and a co-defendant in copyright litigation alongside it; Google's control of search, cloud, and a frontier lab simultaneously , raises competition-law questions several jurisdictions have begun investigating (including China's ongoing antitrust scrutiny of NVIDIA's compliance with Mellanox-acquisition commitments) but none has yet resolved through enforcement action with material effect on market structure.

14.4 Enforcement: Auditing Black-Box Models

The Foundation Model Transparency Index's decline from 58 to 40 (Section 11) is the starkest evidence that voluntary disclosure is not converging toward auditability as capability increases; if anything, the most capable systems are now the least transparent, since leading labs have concluded that disclosing training data, compute, and methodology carries competitive and (they argue) security costs. This directly undermines both regulatory enforcement (the EU AI Office cannot fully verify GPAI providers' systemic-risk self-assessments without disclosed compute and evaluation data) and independent safety research (external auditors cannot red-team or replicate risk assessments for models whose architecture and training data are undisclosed). No jurisdiction has yet mandated third-party model audits with the legal force and technical specificity of, for example, financial-sector audit requirements, though the EU AI Act's conformity-assessment regime for high-risk systems is the closest existing analogue , and its toughest provisions are precisely the ones just delayed to December 2027-August 2028.

15. Possible Solutions

The following policy approaches are presented as a non-exhaustive menu reflecting genuine positions advanced in the current debate; this Background Guide takes no position on which, if any, the Committee should adopt, and delegates should weigh each against their assigned stakeholder's documented interests.

15.1 An International Verification Body ("CERN for AI")

Proposal: establish a standing, internationally staffed body , modelled variously on CERN (shared scientific infrastructure), the IAEA (safeguards and verification), or the IPCC (synthesis of scientific consensus) , with a mandate to verify frontier training runs against agreed compute or capability thresholds, conduct or coordinate independent model audits, and provide pooled compute access for researchers and lower-income states. Proponents (including some FLI-aligned researchers and elements of the UN HLAB's "coherent effort" recommendation) argue this is the only mechanism that can produce genuinely independent verification rather than self-reported compliance. Sceptics note the IAEA analogy is imperfect (nuclear material is physically trackable in a way compute and model weights are not), that no major AI power has endorsed binding external verification of its frontier training activity, and that China's competing World AI Cooperation Organisation proposal suggests any such body would itself become a site of geopolitical contestation over headquarters, staffing, and mandate.

15.2 Compute Tracking and Reporting Thresholds

Proposal: extend the EU AI Act's GPAI "systemic risk" compute threshold (10^25 FLOP) and voluntary frontier-lab self-reporting (as under the Seoul Frontier AI Safety Commitments) into an internationally harmonised, mandatory reporting regime for training runs above an agreed threshold, potentially administered through the UN's new Scientific Panel on AI or a G7/G20-anchored mechanism. This is more technically tractable than full verification (it relies on self-reporting rather than independent audit) and has some precedent in the OECD's pilot Hiroshima Code of Conduct reporting framework. Limitations: self-reported compute figures are unverifiable without underlying audit rights, and the threshold itself becomes quickly outdated as training efficiency improves (smaller

models achieving frontier-comparable results with far less compute, as DeepSeek-R1 demonstrated in early 2025, undermines compute-threshold-based governance specifically).

15.3 Licensing Regimes for Frontier Model Development

Proposal: require developers to obtain a licence, analogous to a broadcasting or pharmaceutical licence, before training or deploying models above a defined capability threshold, with licence conditions covering safety testing, incident reporting, and possibly liability insurance. This approach has support among some existential-risk-focused advocates and echoes elements of the EU AI Act's conformity-assessment regime for high-risk systems, though the Act licenses specific high-risk use-cases rather than frontier model development as such. Critics (including most frontier labs and libertarian-leaning policymakers, reflected in the Trump administration's deregulatory posture) argue licensing entrenches incumbents (only well-capitalised firms can absorb compliance costs), is difficult to calibrate to a fast-moving capability frontier, and has no precedent for successful international harmonisation given the failure of prior efforts to agree binding technology-control regimes outside the CBRN weapons context.

15.4 Open-Source Mandates or Safe Harbours

Proposal: create either affirmative requirements that publicly funded or sufficiently mature AI research be released as open-weight models (to counter concentration, per Hugging Face and some academic advocates), or, conversely, a regulatory "safe harbour" that reduces liability exposure for open-weight developers provided they meet baseline safety-testing and documentation standards before release (an approach some EU AI Act provisions gesture toward for certain open-source GPAI models). This directly engages the Section 10.2 open-source/proliferation tension and would require the Committee to specify a capability threshold above which open-weighting itself triggers additional obligations , a determination on which frontier labs, open-source advocates, and national-security-focused delegates are unlikely to converge without significant negotiation.

15.5 Global Compute and Capacity-Sharing Mechanisms

Proposal: establish pooled or subsidised compute access, cloud credits, and technical capacity-building for researchers and public institutions in low- and middle-income countries, building on the New Delhi Declaration's "AI Commons" and "shared compute infrastructure" working-group deliverables and the UN HLAB's proposed global fund to close compute and data gaps. This directly

targets the Section 10.4 equity gap but requires major compute-holding states and firms to accept some redistribution of a scarce, commercially valuable resource, and existing pledges (the New Delhi "Charter for the Democratic Diffusion of AI") remain voluntary and non-binding, with no enforcement or accountability mechanism specified.

15.6 Digital Public Infrastructure as an Equity Strategy

Proposal: rather than redistributing frontier-model compute directly, prioritise building interoperable, publicly governed digital identity, payments, and data-consent infrastructure (the Indian DPI model championed by Nandan Nilekani) that lets states capture AI-enabled productivity gains in health, finance, and government services without needing sovereign frontier-model capacity. Proponents argue this is more achievable at scale and already has a working template (India's Aadhaar/UPI stack, reportedly informing DPI efforts in an estimated twenty-plus countries). Critics note DPI addresses AI's downstream application layer rather than the underlying compute and model-development concentration, and that population-scale digital-identity infrastructure carries its own surveillance and exclusion risks that require independent safeguards.

16. Questions a Delegate Should Answer

- Does your stakeholder support binding international regulation of frontier AI, voluntary/soft-law coordination, or primarily domestic/unilateral policy , and how would they justify that position given their state's or organisation's documented interests?
- What is your stakeholder's position on compute governance: should training runs above a defined threshold require reporting, licensing, or independent verification, and who should administer any such regime?
- How should the Committee's outcome document treat the EU AI Act's Digital Omnibus delay , as a pragmatic adjustment, a worrying precedent for regulatory retreat, or an internal EU matter outside the Committee's purview?
- Where does your stakeholder stand on open-weight model release: a net democratising force, an unmanageable proliferation risk, or something requiring capability-dependent treatment?
- What concrete, resourced mechanism (not merely aspirational language) would your stakeholder accept to narrow the global compute and AI-investment gap identified in Section 11 , compute-sharing, DPI investment, technology transfer, or something else?
- Should copyright law be reformed specifically for AI training, and if so, toward a licensing/compensation model, an expanded fair-use/text-and-data-mining exception, or continued case-by-case litigation?
- What institutional home, if any, should long-term global AI governance have , an expanded UN mechanism, a G7/G20-anchored body, a new treaty organisation, or continued reliance on parallel national/regional regimes , and how would your stakeholder's influence change under each model?
- How should the Committee balance the near-universal endorsement achieved by voluntary, non-binding declarations (like New Delhi's 89 signatories) against the comparatively narrow but more substantive commitments in binding instruments (like the AI Act) that command far less than universal buy-in?

17. Glossary

Term	Definition
AGI (Artificial General Intelligence)	A hypothesised future system with human-level or broader cognitive ability across most domains, rather than narrow task-specific competence; no consensus technical definition or agreed test for having achieved it currently exists.
ASI (Artificial Superintelligence)	A hypothesised system substantially exceeding human capability across virtually all domains; discussed in long-term risk literature rather than as a near-term engineering target.
Compute	The computational resources used to train or run a model, typically measured in FLOPs; a common proxy for capability and a focus of proposed governance mechanisms.
FLOP	Floating-point operation; the standard unit for measuring the computational work involved in training or running a model.
Parameters / Weights	The numerical values internal to a neural network, learned during training, that determine its behaviour; loosely used as a proxy for model size.
Transformer	The neural network architecture, introduced in 2017, underlying nearly all current large language models, using an "attention" mechanism to weigh relationships between inputs.
Large Language Model (LLM)	A model, typically Transformer-based, trained on large text datasets to predict and generate language.
Foundation Model	A model trained on broad data at scale, intended to be adapted to a wide range of downstream tasks rather than built for one narrow purpose.

Term	Definition
Frontier Model	An informal industry term for the most capable models at the current edge of development; the basis for capability-threshold safety frameworks.
General-Purpose AI (GPAI)	The EU AI Act's regulatory term for a model designed to perform a wide range of tasks, roughly corresponding to "foundation model."
Multimodal Model	A model that can process and/or generate more than one type of data (e.g., text, images, audio) within a single system.
RLHF (Reinforcement Learning from Human Feedback)	A training technique that fine-tunes a model's behaviour using human ratings of its outputs, commonly used to improve helpfulness and instruction-following.
Fine-tuning	Further training of an already-trained model on a narrower dataset to adapt its behaviour for a specific task or style.

18. Chronology

Date	Event
1950	Alan Turing publishes "Computing Machinery and Intelligence," proposing the Turing Test.
1956	The Dartmouth Summer Research Project coins the term "artificial intelligence."
1986	Backpropagation is popularised by Rumelhart, Hinton, and Williams, advancing neural-network training.
1997	IBM's Deep Blue defeats world chess champion Garry Kasparov.
2012	AlexNet's ImageNet win catalyses the deep-learning era.
2016	DeepMind's AlphaGo defeats Lee Sedol at Go.
2017	Google researchers publish "Attention Is All You Need," introducing the Transformer architecture.
2018-2020	OpenAI releases GPT-1, GPT-2, and GPT-3, demonstrating capability gains from scale.
Nov 2021	UNESCO adopts the Recommendation on the Ethics of Artificial Intelligence.
Aug 2022	The US CHIPS and Science Act is signed.
Oct 2022	The US tightens export controls on advanced AI chips to China.
30 Nov 2022	OpenAI releases ChatGPT.
Mar 2023	GPT-4 is released; the FLI "Pause Giant AI Experiments" letter is published; Italy's Garante temporarily bans ChatGPT.
May 2023	Geoffrey Hinton departs Google; the Center for AI Safety extinction-risk statement is published.
Jul 2023	The Frontier Model Forum is founded by Anthropic, Google, Microsoft, and OpenAI.

Date	Event
Aug 2023	China's Interim Measures for the Management of Generative AI Services take effect.
Oct 2023	US Executive Order 14110 is signed; China announces its Global AI Governance Initiative.
Nov 2023	The Bletchley Park AI Safety Summit produces the Bletchley Declaration.
Dec 2023	New York Times Co. v. Microsoft/OpenAI is filed; the UN High-Level Advisory Body publishes its interim report.
Mar 2024	The UN General Assembly adopts a consensus resolution on safe, secure, and trustworthy AI.
May 2024	The Seoul AI Summit produces the Seoul Declaration and Frontier AI Safety Commitments; Colorado's AI Act is signed.
Aug 2024	The EU AI Act enters into force.
Sep 2024	The UN Summit of the Future adopts the Global Digital Compact; the High-Level Advisory Body's final report is published.
Sep 2024	California's Governor vetoes the frontier-model safety bill SB 1047.
Oct 2024	Demis Hassabis is awarded the Nobel Prize in Chemistry; Geoffrey Hinton is awarded the Nobel Prize in Physics.
Dec 2024	South Korea passes its AI Basic Act.
Jan 2025	The Trump administration rescinds Executive Order 14110; DeepSeek releases its R1 model, intensifying the export-control effectiveness debate.
Feb 2025	The Paris AI Action Summit produces a declaration signed by 60+ states, not including the US or UK; the Thomson Reuters v. Ross Intelligence fair-use ruling is issued.

19. Further Reading

Primary Legal and Policy Texts

- Regulation (EU) 2024/1689 (the EU AI Act)
- Interim Measures for the Management of Generative AI Services: Cyberspace Administration of China
- Executive Order 14110 and subsequent 2025 executive orders on AI: US Federal Register

Multilateral Declarations and UN Reports

- The Bletchley Declaration (2023): UK Government summit documentation
- The Seoul Declaration and Frontier AI Safety Commitments (2024): Republic of Korea / UK summit documentation
- The G7 Hiroshima Process International Guiding Principles and Code of Conduct (2023)
- The OECD AI Principles (2019, updated 2024): OECD AI Policy Observatory
- "Governing AI for Humanity": interim (2023) and final (2024) reports of the UN High-Level Advisory Body on AI
- The Global Digital Compact (2024): annex to the Pact for the Future, United Nations

Corporate Governance Frameworks

- Anthropic's Responsible Scaling Policy
- OpenAI's Preparedness Framework
- Google DeepMind's Frontier Safety Framework

Research and Data Sources

- Stanford Institute for Human-Centered AI (HAI), AI Index Report
- OECD AI Policy Observatory
- Epoch AI
- AI Now Institute

Academic Venues

- Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)
- Policy tracks of NeurIPS, ICML, and AAAI
- Harvard Journal of Law & Technology; Yale Journal of Law and Technology