Draft Articles on the Safe Development and Diffusion of Artificial Intelligence

future of life INSTITUTE

The Draft Articles on the Safe Development and Diffusion of Artificial Intelligence are a collaboration between international AI governance experts from multiple organizations, led by the *Future of Life Institute* and the *Centre for International Governance Innovation*. They are designed on the basis of seven building blocks for international AI governance: red lines, safety standards, technical cooperation, emergency response, benefit sharing, institutional arrangements and compliance mechanisms.

For more on these building blocks and a comparative analysis of current efforts on global AI governance, we invite you to visit **global-governance.ai**.

We understand different governments, organizations or individuals have their own views on what an international treaty for AI should include. Therefore, we have designed this as a modular proposal, allowing users to tailor it to their own preferences or priorities. States are ultimately the only actors capable of leading efforts on an international agreement on advanced AI and these proposals are merely food for thought.

To build your own AI treaty, scan the QR code below.

Visit Global Governance of AI website

Table of Contents

Draft Articles on the Safe Development and Diffusion of Artificial Intelligence

José Jaime Villalobos, Duncan Cass-Beggs, Guillem Bas, Mark Brakel, Siméon Campos, Hamza Chaudhry, Matthew da Mota, Oscar Delaney, Talita Dias, Ben Harack, Niki Iliadis, Kevin Kohler, Matthijs Maas, Richard Mallah, Andrew Mazibrada, Adrian Mak, Julia Morse, Sumaya Nur Adan, Henry Papadatos, James Petrie, Charbel Segerie, Saad Siddiqui, Lara Thurnherr, Chloé Touzet, Joanna Wiaterek, Anna Yelizarova[I] [II]

The States Parties,

Recognizing the shared aspiration to foster human well-being, peace, and prosperity through the development and diffusion of technology;

Affirming the positive transformative potential of advanced artificial intelligence ('AI') systems and related advancements in areas like scientific progress, industry and productivity, health, education, governance, as well as reduced inequality and poverty, while also acknowledging that the capacity to develop advanced AI is concentrated in a small number of states;

Noting that certain activities within the lifecycle of potentially autonomous, highly capable, and functionally general purpose advanced artificial intelligence systems, especially their improper or malicious design, development, deployment and use, such as without adequate safeguards, may simultaneously undermine these aspirations, pose national and global security risks, and cause significant harm;

Emphasizing the need for broad and equitable access to AI and its benefits, in a manner that empowers all States, as well as the inalienable right of States Parties to develop, research, produce and use safe AI systems for peaceful purposes without discrimination;

Aware of relevant efforts to advance international understanding and co-operation on AI by multiple States, international and supranational organizations, and other fora; and desiring to support, enhance and complement these existing international arrangements for the governance of advanced AI systems;

Have agreed as follows:

I Listed authors are those who contributed substantive ideas and/or work to this proposal. Contributions include writing, research, and/or review for one or more sections; some authors also contributed content via participation in a May 2025 workshop and/or via subsequent workshops and discussions. As such, with the exception of the lead author, inclusion as author does not imply endorsement of all aspects of these Draft Articles.

II The authors would like to thank Alex Amadori, Adrian Brown, Aaron Scher and Robert Whitfield for insightful comments and feedback. Inclusion in these acknowledgments does not imply endorsement of these Draft Articles.

Article 1. *Object and Purpose*

The present articles are designed to:

1. Ensure that AI benefits all of humanity;
2. Establish conditions for the development of advanced AI systems, to guarantee that they are safe ahead of their deployment;
3. Enumerate adaptable, effective and internationally interoperable safety standards;
4. Facilitate international technical cooperation to advance the objectives of these Articles;
5. Establish international emergency response mechanisms for instances of existing or imminent AI-related emergencies;
6. Establish concrete pathways to distribute the benefits of advanced AI systems to all States Parties and their populations;
7. Establish the institutional arrangements required to achieve these objectives.

Article 2. *Use of terms*

AI provider: a natural or legal person, public authority, agency or other body that develops, deploys or distributes an AI system or that has an AI system developed and places it on the market or puts the AI system into service under its own name or trademark, whether for payment or free of charge.

AI gains: the global profit derived from the development, deployment, distribution or commercialization of an AI system.

AI system: a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments.

AI hardware: any form of hardware, including but not limited to graphic processing units and tensor processing units, employed in the execution of advanced AI computations.

Advanced AI: AI systems that pose systemic or global catastrophic risks.

Systemic risk: a risk that, due to its reach, or due to actual or reasonably foreseeable negative effects on public health, safety, public security, human rights, or the society as a whole, can be propagated at scale.

Global catastrophic risk: a foreseeable and material risk that an AI provider's development, storage, use, or deployment of an AI system will materially contribute to the death of, or serious injury to, more than 50 people or more than one billion dollars (USD $1,000,000,000) in damage to, or loss of, property arising from a single incident involving an AI system.

PART I
Safe Development and Diffusion of AI

Article 3. *Conditions for the development and deployment of AI systems*

1. States Parties retain their sovereign right to pursue, or allow AI providers within their jurisdiction to pursue, the development and deployment of any type of AI system, as long as the safety standards defined under the Articles are implemented.
2. Under no circumstances, are States Parties, or AI providers within their jurisdiction or control, permitted to develop or deploy AI systems that present or may foreseeably present, regardless of the original intent of its providers, any of the uses, risks or capabilities listed in Annex A.
3. Advanced AI providers shall provide sufficient guarantees that an AI system does not present the uses, risks or capabilities listed in Annex A to the International Agency for the Safe Development of AI ('the Agency'). The Agency shall determine whether said guarantees are sufficient or whether additional guarantees are required before the AI system in question can be developed or deployed without the activation of the enforcement mechanism described in Annex C.
4. Following such a determination by the Agency, there will be a continuing duty of review throughout the lifecycle of the AI system concerned whereby AI providers will regularly provide guarantees and the Agency shall determine whether the guarantees provided remain sufficient to ensure the AI system does not present the uses, risks or capabilities listed in Annex A.

Article 4. *Conditions for the development of AI hardware*

1. States Parties shall take all necessary legal, administrative and other measures to make any AI hardware produced within their jurisdiction compatible with safety and security requirements. This includes guaranteeing that any hardware used for advanced AI development has the properties required to operationalize the verification or enforcement mechanisms under the Articles.

2. States Parties shall take measures, individually and through cooperation, to prevent the proliferation of non-compliant AI hardware.

Article 5. *Verification*

1. States Parties agree to ensure compliance with conditions for the safe development and diffusion of AI through a hardware-enabled verification mechanism that is privacy-preserving and secure, described in Annex B.
2. The Agency shall implement the verification mechanism established in Annex B.
3. The Agency may also seek approval by the Conference of States Parties ('the Conference') to implement any other verification mechanisms it deems necessary to reinforce the verification mechanism established in Annex B.

Article 6. *Enforcement of conditions*

1. When the Agency identifies an AI system that enables or pursues any of the prohibited uses, risks or capabilities listed in Annex A, it shall conduct an objective investigation in consultation with the relevant AI provider and, if required, initiate the enforcement mechanism described in Annex C.
2. The Agency may initiate the enforcement mechanism prior to conducting its investigation if the suspected breach poses a global catastrophic risk.
3. The sanctioned AI provider may provide evidence of compliance and request the Agency to lift the enforcement mechanism. The burden will then revert to the Agency to demonstrate continued breach. If the Agency cannot demonstrate continued breach, it shall lift the enforcement mechanism.

Article 7. *Safety standards*

1. Advanced AI providers shall implement standards on the following topics:
    a. Risk governance;
    b. Safety culture and emergency preparedness;
    c. Continuous monitoring and incident reporting;
    d. Information security;
    e. Risk management, including risk identification, risk analysis and evaluation, and system evaluation;
    f. Deployment procedures.
2. The standards under this list shall be developed by a panel of experts appointed by the Agency, in consultation with States Parties, industry and civil society, and in accord-

ance with the best available science. States Parties may nominate experts to the panel and the standards will be subject to a periodic review process during which time the Agency shall invite contributions from States Parties, industry and civil society.

3. The standards developed by the panel of experts shall be approved by the Conference.
4. Advanced AI providers will require a certificate of compliance with safety standards to operate. This certificate will be issued and renewed by a third party evaluator every six months, at the expense of the advanced AI provider being evaluated and in accordance with rates fixed by the Agency.
5. States Parties will implement measures to ensure that advanced AI providers that do not have a certificate of compliance are not able to develop, test, produce, manufacture, otherwise transfer or acquire, or commercialize advanced AI systems.
6. Third party evaluators must submit an evaluation report to the Agency justifying their decision to issue a certificate of compliance.
7. Third party evaluators must have a licence from the Agency to issue certificates of compliance with safety standards. Such a licence shall not be issued except on receipt by the Agency of authenticated documents on the following:
    a. Evidence of relevant expertise;
    b. A declaration of conflicts of interest;
    c. A background check on employees conducting evaluations, which, at a minimum, addresses information security;
    d. Formal submission by the third party evaluator to continuous monitoring by the Agency of its capacity to fulfill the terms of its licence;
    e. Any other documentation reasonably deemed by the Agency to be necessary for the issuing of the licence.
8. The Agency shall establish guidelines for third party evaluators, through a process open to consultation with States Parties, industry and civil society.
9. Evaluation reports and all evidence submitted by third party evaluators to receive a licence shall be published on the website of the Agency, with due regard to sensitive information compromising the national security of State Parties and intellectual property of AI providers.
10. States Parties may challenge certificates of compliance with safety standards by third party evaluators, as well as licences to third party evaluators. A challenge will be submitted to the Agency's Dispute Settlement Panel, which shall decide to accept or reject it within a week.
11. If the Dispute Settlement Panel deems there is a reasonable likelihood that a challenge will succeed, it may suspend a certificate of compliance or licence until the matter has been resolved, or adopt any other provisional measures as it deems appropriate in the circumstances.

12. Interested third parties will be allowed to submit, in writing or, as appropriate, at a public hearing set by the Dispute Settlement Panel, any comments, information, analyses or opinions that they consider relevant to the subject matter of the decision.

Article 8. *Emergency prevention, preparedness and response*

1. States Parties must implement appropriate measures to prevent, prepare for and respond to an AI-related emergency, including **mandatory incident reporting mechanisms by advanced AI providers and** measures to deactivate any AI systems associated with an emergency.
2. States Parties will submit an annual report listing and describing the measures they have implemented to prevent, prepare for and respond to any projected AI-related emergency.
3. Annual reports will be submitted to a peer-review process, through which States Parties shall review and evaluate each other's emergency prevention, preparedness and response capabilities, compliance with emergency preparedness standards, and implementation measures, based on guidelines published by the Agency.
4. If a State Party is determined through the peer-review process to not have implemented appropriate measures to prevent, prepare for or respond to an AI-related emergency, the Agency shall work with the State Party to create an improvement plan detailing corrective actions and timelines for the reviewed State Party.
5. A State Party subject to an improvement plan may request additional resources and assistance from the Agency to facilitate implementation of the required corrective actions.
6. For the purposes of this article, the Agency shall establish protocols for emergency information-sharing, notification and consultation, and coordination among States Parties. Protocols shall also include appropriate information sharing with relevant international entities and the general public, which includes relevant information relating to the hazardous activity, the risk involved and the harm which might result and ascertain their views. Protocols shall also ensure communication is timely, secure, and efficient during emergencies.
7. Without prejudice to any other duty or obligation that it may have under international law, a State Party responsible for harm during an AI-related emergency shall have an obligation to provide adequate assistance and compensation to affected States Parties, for the purpose of victim redress.
8. All other States Parties in a position to do so are also encouraged to provide technical, material and financial assistance to States Parties affected by an AI-related emergency.

PART II
Access and Benefit Distribution

Article 9. *Technical cooperation*

1. Each State Party shall cooperate with other States Parties to achieve the objectives of the Articles.
2. Measures for technical cooperation, coordinated through national or regional contact points, may include:
    a. Joint research and development of technical mechanisms that facilitate the implementation of the Articles, including evaluation, verification and enforcement mechanisms;
    b. Capacity building programs to foster technological self-determination;
    c. Transfer of compute, data and other inputs required for safe AI development;
    d. Mechanisms for the transfer of AI-related benefits within and across jurisdictions;
    e. Joint enterprises to develop AI systems with narrow applications for the public good.

Article 10. *Global AI Benefit-Sharing Framework*

1. States Parties hereby establish a global AI benefit-sharing framework to distribute AI-related benefits to individuals and communities within the jurisdiction of States Parties.
2. The benefit-sharing framework may include, among other distribution mechanisms, a dividend system to deliver recurring payments to residents of States Parties. In that case, each State Party will determine whether the individuals within their jurisdiction receive payments directly from the Global AI Fund or indirectly through national agencies.
3. The global AI benefit-sharing framework will be administered by the Global AI Fund.
4. States Parties shall finance the Global AI Fund through a tax on AI gains of at least five percent, which they will then transfer to the Fund.
5. Advanced AI providers that evade payment of their tax on AI gains will not be eligible to receive a certificate of compliance.

PART III
Institutional Arrangements

Article 11. *National implementation*

Each State Party shall take all necessary legal, administrative and other measures to implement its obligations under the Articles, including:

1. Imposing penal and administrative sanctions, to prevent and suppress any activity prohibited to a State Party under the Articles undertaken by persons under its jurisdiction or control.
2. Establishing national or regional points of contact to coordinate any actions with other States Parties that are required to implement the provisions under the Articles.

Article 12. *Conference of States Parties*

1. The Conference of States Parties shall be composed of all States Parties to the Articles. Each member shall have one representative in the Conference, who may be accompanied by alternates and advisers.
2. The Conference shall meet once per quarter in order to consider and, where necessary, take decisions in respect of any matter with regard to the application or implementation of the Articles, in accordance with its relevant provisions, and on further measures for the attainment of the Articles' objectives.
3. The Conference shall adopt its rules of procedure at its first session, which will take place no later than fifteen days after the entry into force of the Articles.
4. Extraordinary Conferences shall be convened, as may be deemed necessary, by the Agency at the written request of any State Party, provided that this request is supported by at least one third of the States Parties.
5. Extraordinary Conferences may take place through remote secure online participation, which shall be coordinated by the Agency.
6. The Conference shall:
    a. Appoint the Director-General of the Agency;

b. Appoint the members of the Board of the Fund;
c. Approve the rules of procedure of the Agency, submitted by the Agency;
d. Take the necessary measures to ensure compliance with the Articles and to redress and remedy any situation which contravenes the provisions of the Articles;
e. Undertake reviews of the operation of the Articles, taking into account any relevant scientific and technological developments.

7. The Conference shall make decisions by a simple majority, with each State Party having one vote. For some matters, each State Party's voting share shall be calculated on the basis of their contributions to the Global AI Fund, those matters being:

a. The appointment of the Director-General of the Agency;
b. The approval of the rules of procedure of the Agency;
c. The amendment of the verification or enforcement mechanisms established under Annexes B and C, or the approval of additional verification or enforcement mechanisms.

Article 13. *The International Agency for the Safe Development of AI*

1. The States Parties hereby establish the International Agency for the Safe Development of AI.
2. A Director-General shall act as the Agency's head and chief administrative officer.
3. The Director-General shall be appointed by the Conference for a term of three years, renewable for one further term, but not thereafter.
4. The Director-General shall be responsible to the Conference for the appointment of the staff and the organization and functioning of the Agency. The paramount consideration in the employment of the staff and in the determination of the conditions of service shall be the necessity of securing the highest standards of efficiency, competence and integrity. Only citizens of States Parties shall serve as the Director-General or as other members of the staff. Due regard shall be paid to the importance of recruiting the staff on as wide a geographical basis as possible.
5. In the performance of their duties, the Director-General and the other members of the staff shall not seek or receive instructions from any government or from any other source external to the Agency. They shall refrain from any action that might reflect negatively on their positions as international officers responsible only to the Conference.
6. The Agency shall carry out the powers and functions entrusted to it under the Articles, as well as those functions delegated to it by the Conference. In so doing, it shall act in conformity with the recommendations, decisions and guidelines of the Conference and assure their proper and continuous implementation.

7. Each State Party shall respect the exclusively international character of the responsibilities of the Director-General and the other members of the staff and not seek to influence them in the discharge of their responsibilities.
8. The Agency shall promote the effective implementation of, and compliance with, the Articles. It shall cooperate with the national contact point of each State Party and facilitate consultations and cooperation among States Parties at their request.
9. The Agency shall:
    a. Consider and submit to the Conference the draft rules of procedure for the Agency;
    b. Consider and submit to the Conference the draft program and budget of the Agency;
    c. Establish such subsidiary organs as it finds necessary for the exercise of its functions under the Articles, including a Secretariat;
    d. Consider and submit to the Conference the draft report of the Agency on the implementation of the Articles, the report on the performance of its own activities and such special reports as it deems necessary or which the Conference may request;
    e. Make arrangements for the sessions of the Conference including the preparation of the draft agenda;
    f. Conclude agreements or arrangements with States and international organizations, subject to prior approval by the Conference;
    g. Consider guarantees submitted by AI providers that an AI system does not present the risks or capabilities listed in Annex A and decide whether said guarantees are sufficient or additional guarantees are required.
    h. Consider, approve and regularly review the safety standards recommended by a panel of experts appointed by the Director-General;
    i. Issue licenses for third party evaluators;
    j. Publish guidelines for third party evaluators;
    k. Implement the verification and enforcement mechanisms defined under the Articles.
    l. Conduct an annual review of the list of prohibited uses, risks and capabilities in Annex A, as well as the verification and enforcement mechanisms described in Annexes B and C;
    m. Create improvement plans for AI-related emergency prevention, preparedness and response, when required;
    n. Implement, manage, and maintain protocols related to AI-related emergency information-sharing;
    o. Create and maintain arrangements for the involvement of civil society in the processes and decisions of the Agency and the Conference.

10. The Agency will be funded by one percent of States Parties' contributions under the global AI benefit-sharing framework.
11. The Agency shall enjoy on the territory and in any other place under the jurisdiction or control of a State Party such legal capacity and such privileges and immunities as are necessary for the exercise of its functions. The officials of the Agency will similarly enjoy such privileges and immunities as are necessary for the independent exercise of their official functions in connection with the Agency.

Article 14. *The Global AI Fund*

1. The States Parties hereby establish the Global AI Fund to administer the global AI benefit-sharing framework.
2. The Fund will be accountable to and function under the guidance of the Conference of States Parties.
3. The Fund will be governed and supervised by a Board that will have full responsibility for funding decisions.
4. The Fund will establish a secretariat, which will be fully independent. The secretariat will service and be accountable to the Board. It will have effective management capabilities to execute the day-to-day operations of the Fund. The secretariat will be headed by an Executive Director with the necessary experience and skills, who will be appointed by and be accountable to the Board.
5. The Fund shall carry out the powers and functions entrusted to it under the Articles, as well as those functions delegated to it by the Conference. In so doing, it shall act in conformity with the recommendations, decisions and guidelines of the Conference and assure their proper and continuous implementation.
6. The Fund shall:
    a. Receive contributions from States Parties;
    b. Establish the mechanisms for distribution of AI-related benefits, which may include a system for the payment of dividends to residents of States Parties;
    c. Transfer one percent of States Parties' contributions to the Agency, for the Agency's operations.
7. The Fund shall have a trustee with administrative competence to manage the financial assets of the Fund. The trustee will maintain appropriate financial records and will prepare financial statements and other reports required by the Conference, in accordance with internationally accepted fiduciary standards.
8. There will be periodic independent evaluations of the performance of the Fund in order to provide an objective assessment of its results.

9. In order to operate effectively internationally, the Fund will possess legal personality and will have such legal capacity as is necessary for the exercise of its functions and the protection of its interests.
10. The Fund will enjoy such privileges and immunities as are necessary for the fulfillment of its purposes. The officials of the Fund will similarly enjoy such privileges and immunities as are necessary for the independent exercise of their official functions in connection with the Fund.

PART IV
Miscellaneous

Article 15. *Amendments*

1. At any time after entry into force, any State Party may propose amendments to the Articles. The text of a proposed amendment shall be communicated to the Agency, which shall circulate it to all States Parties.
2. The Conference may agree upon amendments which shall be adopted by a positive vote of a majority of two thirds of the States Parties.
3. On the basis of its annual review, the Agency may issue recommendations to the Conference for amendments to the Annexes. Recommendations related to Annex A shall not be made with the intent of removing any of the uses, risks or capabilities listed. Recommendations related to Annexes B and C must be exclusively directed at improving verification or enforcement mechanisms in pursuit of the Articles' objectives.

Article 16. *Settlement of disputes*

1. The States Parties hereby establish a Dispute Settlement Panel as an independent body of the Agency.
2. The Panel shall have jurisdiction in disputes referring to:
    a. Disputes between States Parties concerning the interpretation or application of the Articles;
    b. Disputes between a State Party and the Agency, concerning:
        i. Acts or omissions of the Agency or of a State Party alleged to be in violation of the Articles, or
        ii. Acts of the Agency alleged to be in excess of its mandate or a misuse of power;
    c. Any other disputes for which the jurisdiction of the Panel is specifically provided in this Convention.
3. The Panel shall have the power to indicate, if it considers that the circumstances so require, any provisional measures which ought to be taken to preserve the rights of

the parties to the dispute or to protect the human rights of individuals or communities within the jurisdiction or control of States Parties.

4. The Panel shall give advisory opinions at the request of the Conference on legal questions arising from the Articles.
5. This dispute settlement mechanism is without prejudice of any other dispute settlement mechanisms available to the States Parties under international law.

Article 17. *Universality*

Each State Party shall encourage States not party to these Articles to sign, ratify, accept, approve or accede to the Articles, with the goal of universal adherence of all States to the Articles.

Article 18. *Relationship with other rules of international law*

The present Articles are without prejudice to any obligation incurred by States Parties under relevant treaties or rules of customary international law.

Annex A:
List of prohibited uses, risks and capabilities

1. Risks from enabling chemical, biological, radiological, and nuclear (CBRN) attacks or accidents. This includes significantly lowering the barriers to entry for malicious actors, or significantly increasing the scale or severity of potential or actual harm caused by the design, development, acquisition, release, distribution, or use of related weapons or materials.
2. Uncontrolled cyber-offensive agents capable of disrupting critical infrastructure at a scale that could cause mass disruption or casualties.
3. Loss of control, encompassing any capability or condition that places the continued development, propagation, or operation of an AI system beyond the reach of meaningful human direction, correction, or intervention. This category includes, but is not limited to:
    a. Alignment faking, meaning the propensity of an AI system to temporarily behave or produce outputs in line with human intentions and human values, whether those of the AI provider training it or those of society more broadly, despite having different objectives than the ones displayed through its behavior or output;
    b. Autonomous recursive self-improvement, meaning the automated execution of self-directed tasks that contribute, whether incrementally or through critical capability jumps, to advances in an AI system's capabilities in a manner that outpaces or circumvents adequate human oversight;
    c. Self-replication, meaning the capacity of an AI system to create, maintain, or propagate copies or variants of itself, including through replication strategies that adapt to varying technical or operational circumstances so as to resist or evade shutdown;
    d. Non-interpretability of chain-of-thought, meaning a condition in which the intermediate reasoning steps or explanatory processes generated by an AI system are not comprehensible or verifiable by human evaluators to a degree sufficient to assess the system's alignment with applicable safety requirements and the intentions of its provider; and

e. Autonomous sabotage, meaning the ability to carry out acts of sabotage that substantially raise global catastrophic risk, through access to sensitive assets, subterfuge or any other means.

4. Highly effective epistemic manipulation, meaning the systematic subversion of individuals' epistemic autonomy through the deliberate or foreseeable use of techniques that bypass rational deliberation to materially distort political, electoral, or civic decision-making.

Annex B.1:
[Existing] Verification mechanisms

1. To verify States Parties' obligations under Articles 3 and 4, the Agency may take any of the following measures, independently or in combination:
    a. Interview personnel of advanced AI providers or AI hardware manufacturers, with interviews having a narrow scope defined by the Conference;
    b. Intelligence gathering activities, including data and code validation, financial audits, sanity checks, workload classification, analytic output verification, satellite or aerial imagery, measurements of energy consumption, thermal and networking characteristics, open-source information, public and privately disclosed statements from data centers, recordings from cameras at AI data centers, and other forms of human intelligence, signals intelligence or electronic surveillance. The Agency shall disclose these activities to the States Parties in which the activities are being carried out, and at all times ensure that any activity is necessary and proportionate to its objectives, as well as compliant with international human rights law.
    c. Maintaining a chip registry that includes a unique ID and the declared location of chips that are regularly used for advanced AI development or deployment.
    d. Location verification through a network delay-based system in which trusted servers ping the secure ID of chips to measure round-trip latency.
    e. On-site inspections of advanced AI providers' facilities, data centers or hardware manufacturing facilities. These inspections will be carried out in strict compliance with protocols developed by the Agency and approved by the Conference.
2. States Parties shall implement whistleblower programs within their jurisdictions to enable and encourage personnel of advanced AI providers to unveil potential violations of the Articles, to both domestic entities and the Agency.

Annex B.2:
[Hardware-enabled] Verification mechanism

To verify States Parties' obligations under Articles 3 and 4, the Agency will take the measures described below.

The Agency shall install, under close supervision of States Parties, one or more neutral data centers (also referred to in this annex as 'verifier's data centers'). It shall also coordinate with States Parties to repurpose or build one or more provers' data centers. All data centers will contain verified hardware, meaning hardware that has been approved by the Agency as complying with the specifications required to carry out the verification mechanism described in this Annex. Hardware specifications must include the ability to demonstrate that tampering has taken place. Agency officials shall routinely inspect both the verifier and the provers' data centers to ensure they comply with the predetermined specifications. States Parties may also continuously monitor both the verifier and the provers' data centers, following protocols developed by the Agency.

The neutral data centers shall be under the Agency's ownership and control, but with access to States Parties under the aforementioned Agency protocols. The Agency will be directly accountable to the Conference of States Parties for the administration of the neutral data center. The provers' data centers shall be located in the jurisdiction of States Parties which host hardware used to develop advanced AI.

All verified data centers shall have a digital perimeter, meaning a physical, digital, or hybrid boundary system that prevents data from entering or exiting the enclosed area without being noticed by both the verifier and the prover. The specifications of this digital perimeter shall be defined by the Agency in consultation with States Parties.

Within the digital perimeter of their prover's data center, AI providers shall create cryptographic commitments, meaning the short strings that commit to a piece of data without revealing the actual data, thus allowing the prover to later reveal that data and simultaneously prove that it was the same data that was committed to earlier. The Agency shall also create cryptographic commitments on the computations it has used to verify that an AI system complies with the Articles. The Agency shall publish the results of verification in a way that does not compromise the provers' data.

Annex C.1:
[Domestic] Enforcement mechanism

When the Agency identifies an AI system that pursues or is capable of any of the prohibited uses defined in article 3, or presents any of the risks or capabilities listed in Annex A, it shall notify any States Parties with jurisdiction over the violation. These States Parties shall execute the enforcement mechanism described below.

All chips used for AI development or deployment are subject to an authorization to operate, which is given by the State Party in which an AI provider is located. This authorization will be automatically renewed periodically (potentially every few minutes) unless a violation of article 3 is verified by the Agency. The authorization will consist of a certificate that is cryptographically signed by the State Party's private key. Firmware running on the chip's security processor will verify that the signature is from a private key that coincides with the State Party's public key, also stored in the firmware binary. The identification number of each cryptographic signature will be securely stored in the chip once it has received a new authorization. If a new authorization does not meet these conditions, the chip will refuse to complete any new computations until the State Party provides a new valid authorization.

If a violation of article 3 is verified by the Agency, the relevant State Party shall not extend a new authorization for the chips being used to run the AI system found to be in violation. At the start of the next clock cycle, those chips will detect a lack of authorization and refuse to complete any computations until the State Party provides a new authorization.

All chips used for AI development must be placed in a tamper-proof secure enclosure along with an auxiliary guarantee processor. All of the chip's memory and other components will also be contained within the tamper-proof secure enclosure. If an attempt is made to tamper with the enclosure, the chip will automatically wipe the guarantee processor's firmware and microscopic fuses on the AI chip will be activated to render the chip permanently inoperable.

States Parties shall install a firmware update in guarantee processors as regularly as needed to improve the processor's security or implement changes in the conditions for authorizations.

Annex C.2:
[International] Enforcement mechanism

When the Agency identifies an AI system that pursues or is capable of any of the prohibited uses defined in article 3, or presents any of the risks or capabilities listed in Annex A, it shall execute the enforcement mechanisms described below.

All chips used for AI development or deployment are subject to an authorization to operate, which is given by the Agency. This authorization will be automatically renewed periodically (potentially every few minutes) unless a violation of article 3 is verified by the Agency. The authorization will consist of a certificate that is cryptographically signed by the Agency's private key. Firmware running on the chip's security processor will verify that the signature is from a private key that coincides with the Agency's public key, also stored in the firmware binary. The identification number of each cryptographic signature will be securely stored in the chip once it has received a new authorization. If a new authorization does not meet these conditions, the chip will refuse to complete any new computations until the Agency provides a new valid authorization.

If a violation of article 3 is verified by the Agency, it will not extend a new authorization for the chips being used to run the AI system found to be in violation. At the start of the next clock cycle, those chips will detect a lack of authorization and refuse to complete any computations until the Agency provides a new authorization.

All chips used for AI development must be placed in a tamper-proof secure enclosure along with an auxiliary guarantee processor. All of the chip's memory and other components will also be contained within the tamper-proof secure enclosure. If an attempt is made to tamper with the enclosure, the chip will automatically wipe the guarantee processor's firmware and microscopic fuses on the AI chip will be activated to render the chip permanently inoperable.

The Agency may require States Parties to install a firmware update in guarantee processors as regularly as needed to improve the processor's security or implement changes in the conditions for authorizations.

Commentaries

General commentary

(1) Recognition of the need for legitimate and effective international cooperation on the governance of artificial intelligence (AI) is likely to keep growing in the coming months and years as the capabilities of AI systems expand at an exceptionally fast pace.[1] AI is having transformational global implications,[2] with harms from this technology already materializing[3] and many others starting to appear over the horizon.[4] Awareness around these risks is emerging gradually as AI systems continue to exhibit dangerous capabilities, and could increase as a result of a sudden "risk awareness moment" or "focusing event".[5] As stated by the UN High-Level Advisory Body on Artificial Intelligence, "waiting for a threat to emerge may mean that any response will come too late".[6] The international community still has time to prevent what could very well become a wave of individual, collective and State harm, but it will need to act decisively to create the norms and institutions required to do so.

(2) Under these circumstances, decision makers are likely to seek guidance from experts on specific mechanisms, institutional proposals, and language to include in international instruments that address global AI challenges. This type of guidance is currently rare,[7] so there is little for governments and other stakeholders to borrow from when the time comes.

(3) There are historical examples of proposals from civil society having a significant influence on foundational international instruments. For instance, the Universal Declaration on Human Rights borrowed extensively from the American Law Institute's Statement of Essential Rights ('ALI Statement'),[8] and the Anti-Ballistic Missile Treaty was largely shaped by the epistemic community in the United States and its exchanges with the Soviet epistemic community.[9] This document–a concerted proposal by reputable experts and organizations with experience in international AI governance–seeks to fill that role concerning the international governance of AI.

(4) The Draft Articles on the Safe Development and Diffusion of Artificial Intelligence ('the Articles') build upon existing cross-cutting international obligations that are not specific to AI, as well as valuable, but insufficient, efforts to govern AI. It is indeed valuable to acknowledge that a vast array of international obligations already apply to the development of AI and its impacts. An overview of these can be found in Villalobos, Maas and Winter (2026), drawing from international disaster law, international human rights law, international criminal and humanitarian law, international environmental and economic law, and emerging future generations law.[10] At the same time, the particular issues presented by advanced AI may lead to interpretative gray zones or gaps that are best filled by a bespoke international instrument. In this regard, to date, there are more than two

dozen international initiatives to address global AI challenges–ranging from resolutions from the United Nations General Assembly to national, minilateral and regional statements or instruments.[11] While these initiatives represent meaningful progress, they have yet to fully meet several widely recognized criteria for an effective international AI governance framework. In particular, gaps remain with respect to: (i) binding legal obligations; (ii) direct relevance to risks, including large-scale risks; (iii) a pathway towards the international institutions required to enforce a binding agreement; and (iv) the inclusion of every State, both regarding participation in the formulation of the regime and in access to the regime's benefits.

(5) Voluntary commitments, high-level principles and other forms of soft law play an important part in any international governance regime. They shape norms, affect State and industry behavior, and in many cases form the basis for hard law.[12] Initiatives like the OECD's AI Principles,[13] the Bletchley[14] and Seoul[15] Declarations, the UNESCO Recommendation on the Ethics of AI,[16] and the G7's Hiroshima AI Process[17] have made an invaluable contribution in this regard. Following the natural evolution of international law, however, these soft law instruments should now give way to hard law. **Binding global rules** are conducive to setting an interoperable framework for AI marketization, giving the AI industry legal certainty in their transnational operations.[18] At the same time, a binding legal framework is crucial to ensure that the industry's interests are not placed ahead of users' wellbeing. Voluntary commitments cannot be enforced by States or international organizations, so they rely on the good faith of corporations with (often legally enforceable) profit incentives that may be incompatible with or trump those same corporations' efforts to have a positive impact in the world. This is the principal motivation behind the application of legal standards and rules to most other commercial activities. It is also the reason why State or private activities with the potential to cause transboundary harm are often subject to international rules. The only existing treaty on AI development, the Council of Europe's Framework Convention on Artificial Intelligence, human rights, democracy and the rule of law, has this intention.[19] Unfortunately, it has not yet entered into force and, even if it did, it has serious limitations. The agreement is centred around well-intentioned but vague obligations that are very open to interpretation and it does not include verification or enforcement mechanisms beyond unilateral transparency measures. Concerningly, article 3(1)(b) gives parties the choice (at signature or ratification) to exclude private activity from the scope of the treaty's obligations, which effectively hollows the convention, considering most AI development is carried out by private corporations. Finally, while the treaty is open to ratification by any country, its negotiation process was not globally representative and is therefore perceived as a Western/Global North initiative.

(6) There are numerous examples in international law of States coming together to address large scale risks. States have reached agreements to limit the negative externalities of, for example, nuclear weapons, biological weapons, pandemics, ozone layer depleting substances, and greenhouse gas emissions that accelerate climate change.[20] In all of these cases, there has been a mutual understanding among States that one or more international agreements are necessary to **directly tackle the risks at hand**. In the case of AI, many of the risks have already been identified and plausible pathways towards their materialization have been traced. The International AI Safety Report groups these risks in three categories: malicious use (e.g., manipulation of public opinion, biological and chemical attacks), malfunctions (e.g., loss of control over AI systems that could start pursuing their own goals), and systemic (e.g., mass job loss, excessive market concen-

tration).[21] Some of these risks have started to materialize,[22] prompting initial reactions from the international community. While nascent international efforts such as the Hiroshima AI Process represent a constructive initial step toward addressing the criterion of direct relevance to risks, they do not yet fully satisfy the other three criteria outlined here.

(7) It has been empirically demonstrated that an international regime without enforcement mechanisms, and the required institutional arrangements to implement those mechanisms, will almost certainly fail.[23] Ensuring that we all benefit from AI while its risks are sufficiently mitigated will require **one or more international institutions that implement and enforce the mechanisms and rules that States have agreed to**. Dozens of institutional designs for AI governance have been recommended by experts, States and international organizations.[24] With the potential exception of the UN Independent International Scientific Panel on AI and the Global Dialogue on AI Governance,[25] these institutions have not yet been created.

(8) Both the upsides and downsides of AI technologies are global. AI promises to increase the living standards of everyone on Earth by contributing to advances in areas like scientific progress, education, health or climate change mitigation.[26] At the same time, risks and harms from AI systems can affect all of us, regardless of whether we have contributed to their development or even used them. In light of this shared impact, **the formulation of an international AI governance regime should be inclusive of all States**. At a minimum, it should include States where the most advanced capabilities have been developed, which are capable of increasing or decreasing risks much more than other States. Without this inclusivity, some States may choose to pursue AI development outside of the parameters defined by the international community–even if it is in the medium to long-term future–[27]but they might also be locked out of the opportunities and benefits that come along with international cooperation on AI. Smaller, more targeted minilateral efforts like those pursued by the G7 or the EU can continue to play an important role particularly given their ability to act with speed and coordination. These efforts may serve as important interim steps or complements, but they need not be seen as substitutes for more comprehensive, multilateral approaches.

(9) On the basis of recent studies on gaps and opportunities for an international AI governance regime,[28] these Articles aim to fulfill the four criteria discussed above through a framework that addresses seven building blocks for an international governance regime. These are: red lines, safety standards, technical cooperation, emergency response, benefit sharing, compliance mechanisms, and institutional design. Each of these building blocks are discussed below in the corresponding sections of the Articles.

(10) Together, the building blocks represented in the Articles aim to ensure that AI technologies can disseminate and benefit humanity, allowing individuals and societies to thrive and meet the goals they have set for themselves. Ultimately, however, States have a duty to ensure that this development or transfer takes place in the only way where those benefits can be truly realized. That is, in the same way that other emerging technologies and most products are treated around the world: in a responsible manner that prioritizes people's safety and wellbeing.

Preamble

The States Parties,

Recognising the shared aspiration to foster human well-being, peace, and prosperity through the development and diffusion of technology;

Affirming the positive transformative potential of advanced artificial intelligence ('AI') systems and related advancements in areas like scientific progress, industry and productivity, health, education, governance, as well as reduced inequality and poverty, while also acknowledging that the capacity to develop advanced AI is concentrated in a small number of states;

Noting that certain activities within the lifecycle of potentially autonomous, highly capable, and functionally general purpose advanced artificial intelligence systems, especially their improper or malicious design, development, deployment and use, such as without adequate safeguards, may simultaneously undermine these aspirations, pose national and global security risks, and cause significant harm;

Emphasizing the need for broad and equitable access to AI and its benefits, in a manner that empowers all States, as well as the inalienable right of States Parties to develop, research, produce and use safe AI systems for peaceful purposes without discrimination;

Aware of relevant efforts to advance international understanding and co-operation on AI by multiple States, international and supranational organizations, and other fora; and desiring to support, enhance and complement these existing international arrangements for the governance of advanced AI systems;

Have agreed as follows:

Article 1. *Object and Purpose*

The present articles are designed to:

1. Ensure that AI benefits all of humanity;
2. Establish conditions for the development of advanced AI systems, to guarantee that they are safe ahead of their deployment;
3. Enumerate adaptable, effective and internationally interoperable safety standards;
4. Facilitate international technical cooperation to advance the objectives of these Articles;
5. Establish international emergency response mechanisms for instances of existing or imminent AI-related emergencies;
6. Establish concrete pathways to distribute the benefits of advanced AI systems to all States Parties and their populations;
7. Establish the institutional arrangements required to achieve these objectives.

(1) The preamble and Article 1 set out the Articles' justification, scope and object. These two sections reiterate the need for international AI governance to achieve a balance between access to the benefits of AI and the mitigation of its risks.[29] The Articles start from the premise that, indeed, these two objectives are not in opposition but are closely linked and complementary to each other. Humanity cannot reap the benefits of AI if it does not first ensure that the technology is trustworthy, safe and secure. Currently, the incentives of AI development are set in such a way that corporations are concentrating power and racing towards the creation of advanced AI at the expense of safety, prioritizing profit-driven objectives rather than AI applications that can serve as a tool to benefit humanity.[30] Therefore, these Articles pursue the **safe diffusion of advanced AI**, so that all of humanity can benefit from the technology without needing to sacrifice its safety, rather than the non-proliferation of advanced AI.

(2) The core content and goals–or to put it in legal terms, the object and purpose– of the Articles were set through an analysis of patterns in international law and an exploration of current gaps in the international governance of AI. Regarding the first of these elements, Villalobos, Maas and Winter (2026) conducts a comparative analysis of the international legal norms applicable to risks similar to those posed by AI, identifying close to sixty rules from both risk-specific regimes (biological weapons, nuclear weapons, pandemics, ozone layer depletion, and extreme climate change) and cross-cutting regimes (international disaster law, international human rights law, international criminal and humanitarian law, international environmental law, international economic law, and the international law of future generations). Among these, the authors unveil seven patterns that can be applied to any source of global catastrophic risk, including advanced AI. These patterns, which may be interpreted as general principles of law, refer to the obligations to contain, prevent, notify, share, sanction, adapt, and weigh interests when designing legal obligations or policy related to large-scale risks.[31]

(3) Concerning gaps in international AI governance, Maas and Villalobos (2023) identify seven models for international institutions to govern advanced AI, discussing each model's function, common examples, unexplored examples, academic and governmental proposals, as well as critiques. The seven models are: scientific consensus-building; political consensus-building and norm-setting; coordination of policy and regulation; enforcement of standards or restrictions; stabilization and emergency response; international joint research; and distribution of benefits or access.[32] Additionally, Cass-Beggs, Clare and others (2024) structure the gaps in international AI governance into three main challenges: how to realize and share the benefits of AI, how to mitigate severe global risks posed by AI, and how to make legitimate and effective decisions about the future implications of AI for humanity.[33]

(4) When merged, these three analyses show there are at least seven building blocks for international AI governance. These building blocks are: red lines, safety standards, technical cooperation, emergency response, benefit sharing, compliance mechanisms, and institutional arrangements.[34] The objects embedded in Article 1 are a reflection of these building blocks and inform the contents of these Articles. The justification and proposals for each one are explained throughout this document in the commentary for each article.

Article 2. *Use of terms*

AI provider: a natural or legal person, public authority, agency or other body that develops, deploys or distributes an AI system or that has an AI system developed and places it on the market or puts the AI system into service under its own name or trademark, whether for payment or free of charge.

AI gains: the global profit derived from the development, deployment, distribution or commercialization of an AI system.

AI system: a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments.

AI hardware: any form of hardware, including but not limited to graphic processing units, employed in the execution of advanced AI computations.

Advanced AI: AI systems that pose systemic or global catastrophic risks.

Systemic risk: a risk that, due to its reach, or due to actual or reasonably foreseeable negative effects on public health, safety, public security, human rights, or the society as a whole, can be propagated at scale.

Global catastrophic risk: a foreseeable and material risk that an AI provider's development, storage, use, or deployment of an AI system will materially contribute to the death of, or serious injury to, more than 50 people or more than one billion dollars (USD$1,000,000,000) in damage to, or loss of, property arising from a single incident involving an AI system.

(1) This clause defines key terms used throughout the Articles. It attempts to borrow from existing legislation to prevent normative conflicts and to facilitate the harmonization of domestic rules on AI with international law, reducing barriers to interpretation and application of the rules in these Articles and making it easier for governments and the AI industry to apply them. Thus, the definitions for "AI provider", "AI system" and "systemic risk" match or are very similar to those found under article 3 of the EU AI Act,[35] the most comprehensive piece of legislation on AI to date. It should be noted, that for the purposes of the Articles, the term 'AI model' is purposefully omitted, as its obligations pertain to the downstream applications that can produce harm. For practical purposes, even if 'model' and 'system' are technically different concepts, the term 'system' should be understood to encompass models and the two are at times used interchangeably throughout this document. The definition for "global catastrophic risk" is an adaptation of the term used under the Transparency in Frontier Artificial Intelligence Act (SB53) from California,[36] the first law in the United States to directly address risks from advanced AI.

(2) Only two definitions under Article 2 are novel, given the terms have not yet been employed by domestic legislation in any country. The first is "AI gains", which acts as the basis for contributions towards the global AI benefit-sharing framework established under articles 10 and 14. The second is "AI hardware", which captures all the physical equipment involved in data-processing by an AI system. This term is the basis for the obligations under article 4 and indirectly acts as the anchor for the hardware-enabled governance mechanisms under articles 5 and 6, reflected in annexes B and C.

PART I
Safe Development and Diffusion of AI

Article 3. *Conditions for the development and deployment of AI systems*

1. States Parties retain the sovereign right to pursue, or allow AI providers within their jurisdiction to pursue, the development and deployment of any type of AI system, as long as the safety standards defined under the Articles are implemented.
2. Under no circumstances, are States Parties, or AI providers within their jurisdiction or control, permitted to develop or deploy AI systems that present or may foreseeably present, regardless of the original intent of its providers, any of the uses, risks or capabilities listed in Annex A.
3. Advanced AI providers shall provide sufficient guarantees that an AI system does not present the uses, risks or capabilities listed in Annex A to the International Agency for the Safe Development of AI ('the Agency'). The Agency shall determine whether said guarantees are sufficient or whether additional guarantees are required before the AI system in question can be developed or deployed without the activation of the enforcement mechanism described in Annex C.
4. Following such a determination by the Agency, there will be a continuing duty of review throughout the lifecycle of the AI system concerned whereby AI providers will regularly provide guarantees and the Agency shall determine whether the guarantees provided remain sufficient to ensure the AI system does not present the uses, risks or capabilities listed in Annex A.

(1) In line with the principle of national sovereignty,[37] States should be able to pursue the development of a technology as transformative as AI for the betterment of their people. Moreover, they should be able to do so in a way that is guided by their national values and that narrowly adapts to their national interests and objectives, their particular challenges, and their own vision of the future. Native AI systems applied to local problems can only enhance technological self-determination and the right to science.[38]

(2) A non-proliferation regime, similar to the one set in place for nuclear weapons, is unlikely to be effective to mitigate risks from AI.[39] There are, to be clear, valid safety-related arguments in favor of these controls that go beyond geostrategic justifications.[40] Namely, the diffusion of advanced chips without appropriate safeguards could lead to the accumulation of compute power by malicious actors willing to misuse AI or could increase the chances of loss of control.[41] However, these controls are ultimately an unsustainable strategy. First, because the amount of compute needed to develop an advanced AI system is likely to diminish with time as training runs become more efficient.[42] Additionally, many countries will, in principle, be able to achieve self-sufficiency for the development of AI, preventing them from having to rely on the international supply of compute or other AI precursors.[43] Targeted non-proliferation efforts could be helpful in creating an adaptation buffer that allows experts to forecast concerning capabilities and develop adequate mitigation measures.[44] However, this strategy does not have to depend on

export controls that prevent good actors from accessing compute. It can rely on an international governance framework that creates effective red lines and safety standards around AI development for *any* actor. These Draft Articles pursue an approach of 'safe diffusion' rather than non-proliferation.

(3) At the same time, national sovereignty and self-determination have limitations,[45] just like the vast majority of other rules and principles of international law have limitations or exceptions.[46] These include, among many others, the no harm principle, the obligation of due diligence, and, relatedly, the respect, promotion and fulfillment of human rights. As such, States' freedom to advance or adopt AI within their jurisdictions must be subjected to necessary and proportional limitations that ensure the technology proliferates–as it should–but in a responsible way that protects users, collectives and States themselves from harm.[47]

(4) Some sources of harm from AI have already been addressed by a number of States.[48] However, many of the largest risks posed by AI remain under-regulated both nationally and internationally even when there are strong reasons for the international community to prohibit certain uses and capabilities associated with those risks (see commentary under Annex A).

(5) Article 3.2 follows the modular framework of other international agreements that define the scope of application for their obligations in appendices or annexes.[49] There are two main purposes for the adoption of this model, both related to the desired flexibility of the Articles' provisions. The first is to be able to include risks and capabilities that emerge as a result of technological progress.[50] This is key in the context of AI, which is still a nascent technology in its most advanced forms and could therefore continue to demonstrate capabilities that it does not currently have or that even experts in the field are not able to predict.[51] Connectedly, the second justification for this modular approach is to facilitate amendments to the annexes, as explained below in the commentary to Article 15. Commentary on the list of uses, risks and capabilities can also be found below under Annex A.

(6) Most of the conditions under paragraph 2 and under Annex A are categorical–that is, they belong to a set category, rather than being determined by underlying continuous variables (e.g., number of casualties or floating operation points).[52] Categorical norms are more resilient to perceptual errors and are more sustainable over time compared to norms based on thresholds or continuous measures. Threshold norms introduce criteria that may result in ambiguity if states disagree on the precise threshold.[53] Thus, categorical conditions like the ones set in Article 3 are more legally resilient–they have a higher capacity to remain effective and authoritative over time, despite technological change.[54] That is not to say that thresholds do not have a role to play in international AI governance. Indeed, some of the elements under Annex A may involve setting thresholds to define the limit between acceptable and unacceptable risks, uses or behaviors. Experts tend to agree that if thresholds meet certain conditions, they have a better chance of eliciting proportional policy responses to certain levels of risk. Among these conditions, thresholds should be set by multiple stakeholders, serve a specific purpose, be verifiable by external actors and account for uncertainties, be tied to a reporting system to notify when they have been exceeded, accommodate uncertainties in risk estimation and mitigation, and identify types of harms and model the severity of their impacts.[55] Therefore, Article 3 does not exclude the use of thresholds by the Agency as a way of determining which guarantees are deemed sufficient to demonstrate that a system does not present the

uses, risks or capabilities defined in Annex A. The process of defining these thresholds, however, is delegated to the implementation phase of the Draft Articles instead of being directly embedded in the language of Article 3 to allow for alternative methodologies.

(7) Article 3.4 deals with guarantees required from AI providers in relation to the list of risks and capabilities in Annex A. These guarantees must be *sufficient* from the perspective of the Agency, meaning that the Agency must be satisfied that the evidence provided by AI providers demonstrates that the AI systems they are developing or deploying do not pose the risks listed in Annex A. For each risk or capability, this may imply different forms of evidence, including reports from third party evaluators, different thresholds, or several tiers or degrees of proof commensurate with potential harm.[56] The burden of proof is placed on the providers, rather than the Agency, given that their proximity to the activities being carried out provides them with the access and expertise required to produce the necessary evidence. This approach also prevents an overly-specific type of mitigation measure or guarantee to be indiscriminately applied to all forms of risk or capability, regardless of whether they are the optimal form of evidence for those. This model follows the practice of other industries, where the State sets a general duty to mitigate risks–e.g., in the form of due diligence requirements or environmental impact assessments–and the provider is required to ideate the measures or safeguards required to prevent harm, at the satisfaction of a State agency or an international organization.[57] The Agency is not precluded from publishing guidelines or frameworks specifying minimum evidentiary standards per risk category, but it should be clear that these are a floor and not a ceiling, and that the Agency would maintain its ability to seek further evidence to reach the threshold of sufficiency established in Article 3.3.

(8) The article does not specify the form by which the Agency will determine whether a guarantee is sufficient or not (see commentary above). It prefers delegating this process to the Agency itself, while noting that–as discussed below under Article 13–this body should be staffed adequately to fulfill this function. The type of approval the Agency must give is also not predetermined,[58] but the Article is clear that if any advanced AI system is deployed without this approval, the Agency will automatically activate the enforcement mechanism under article 6, under the assumption that the system is excessively hazardous. An AI provider would have to provide evidence to the contrary for the enforcement mechanism to be lifted. It is also worth noting that any advanced AI providers who have been granted an approval by the Agency are still subjected to the verification mechanism under Article 5. Moreover, the enforcement mechanism under Article 6 could still be activated if a breach of the conditions set under paragraphs 2 and 3 is detected–regardless of whether the advanced AI provider had obtained approval from the Agency or not.

Article 4. *Conditions for the development of AI hardware*

1. States Parties shall take all necessary legal, administrative and other measures to make any AI hardware produced within their jurisdiction compatible with safety and security requirements. This includes guaranteeing that any hardware used for advanced AI development has the properties required to operationalize the verification or enforcement mechanisms under the Articles.
2. States Parties shall take measures, individually and through cooperation, to prevent the proliferation of non-compliant AI hardware.

(1) While the target of the conditions under article 3 are related to the potential externalities of an AI system, article 4 focuses on AI precursors; namely, AI hardware (for the scope of this term, see the commentary under Article 2). AI hardware–most importantly, chips–is a useful target given that it is detectable, excludable, quantifiable, and its supply chain is concentrated in a small number of actors. Therefore, it can be used, among other things, to build guardrails into hardware to verify compliance with international norms and standards.[59]

(2) The safety and security requirements of hardware under article 4 are moving, rather than stationary, targets. Technological innovation will continue to lead to new and better ways in which AI hardware can be made safer or more secure. Therefore, the first sentence of article 4 establishes an obligation of conduct or of means, rather than an obligation of result. States Parties must be able to demonstrate that they are *taking all necessary measures*[60] to implement techniques that make AI hardware compatible with safety requirements–which includes continuous research and development of new safety mechanisms–but they do not need to reach a predetermined threshold for safety to meet this obligation and they are not limited by a pre-established list of safety properties.[61] The same applies to security requirements–security understood as "the property of being resilient to technical interference, such as cyberattacks or leaks of the underlying model's source code."[62]

(3) The only exception to this norm is found in the second sentence of article 4.1, which does establish an obligation of result. States Parties must guarantee that any AI hardware produced within their jurisdiction has the properties required by the verification and enforcement mechanisms established in annexes B and C. As explored in more detail in the commentary to Article 5, hardware-enabled verification mechanisms can be a more politically feasible alternative to the type of verification mechanisms found in other international governance regimes. Among other reasons, hardware-enabled mechanisms can allow competing actors to verify each others' claims in a way that protects sensitive information on intellectual property or national security. It could be argued, indeed, that an effective international governance regime for AI relies on the availability of this type of verification mechanism. In the case of enforcement, while there is a larger pool of solutions to the enforcement of international obligations relating to AI, the Articles rely on a hardware-enabled enforcement mechanism that requires that hardware to meet certain criteria (for reasons explored under the commentary to Annex C). Thus, the obligation under the latter part of article 4.1 is one of result because it is a *sine qua non* condition for the verification mechanism under Annex B.2 and the enforcement mechanism in Annex C, both required to implement the Articles.

(4) Article 4.2 addresses the risk of allowing advanced AI hardware that cannot be verified to proliferate without adequate guardrails, potentially leading to the accumulation of a large cluster of hardware that can be used to undermine the Articles' objectives. There are numerous ways in which States could carry out this obligation, including chip registries, on-chip controls, off-chip analog sensors and export controls or sanctions coordinated among States Parties.[63]

Article 5. *Verification*

1. States Parties agree to ensure compliance with conditions for the safe development and diffusion of AI through a hardware-enabled verification mechanism that is privacy-preserving and secure, described in Annex B.
2. The Agency shall implement the verification mechanism established in Annex B.
3. The Agency may also seek approval by the Conference of States Parties ('the Conference') to implement any other verification mechanisms it deems necessary to reinforce the verification mechanism established in Annex B.

(1) International agreements often hinge on the credibility of commitments–States often need assurances that others will uphold their end of the bargain. This gave rise to the maxim "trust but verify", often used in Cold War arms control. In practice, verification mechanisms are the tools and processes by which parties gather information to monitor compliance with an agreement. States can be reluctant to enter agreements without reliable ways to produce timely information on compliance and the presence of ongoing verification measures that can determine an agreement's success.[64] Verification can take many forms, from legal-institutional arrangements (e.g., reporting requirements, inspections by international agencies) to technical measures (e.g., satellite surveillance or forensic analysis).[65] Generally, verification reduces the pervasive uncertainty in international relations and builds mutual trust without which cooperation would be very difficult.

(2) One practical challenge for a successful verification mechanism lies in the inherent trade-off between transparency and security. For a prohibition agreement to be viable, verification must be sufficiently transparent to ensure compliance. This might necessitate revealing information about a State's capabilities, which can simultaneously compromise its security by exposing vulnerabilities.[66] It might also expose a company's intellectual property, making it easier for an adversary to mimic a technology or product to compete. Near-term verification mechanisms, such as those included in Annex B.1, are transparent but are less secure and trustworthy than ideal for a sustainable international AI governance regime. While many of the challenges posed by these mechanisms can be overcome through *hardware-enabled verification mechanisms,* including the one described in Annex B.2, these still face technical hurdles that can only be overcome through significant investments in research and development. Because hardware-enabled verification mechanisms can be more privacy-preserving and secure than near-term verification mechanisms, it may be in States' interest to invest early in them. Early investments in these longer-term mechanisms are also important to build confidence among States Parties that a robust international regime with effective verification mechanisms is possible and will not be used to steal intellectual property or information that is sensitive to national security.[67]

(3) The Articles relegate the full description of both the verification and enforcement mechanisms to annexes (B and C). The reason behind this coincides with the justification for the list of risks and capabilities in Annex A. The Articles can lock in general obligations for States Parties or a mandate for the Agency without locking the specific technical properties of the implementation methods in the text of the treaty. This makes it easier for States Parties to make amendments to those methods in line with technological progress, making them more resilient to future needs or demands.

(4) Following the common practice of other international verification mechanisms, the Articles give the mandate of implementing the mechanism in Annex B to the Agency. There

are numerous reasons why an international organization like the Agency is well-placed to carry out this function: among others, it is a politically neutral actor with no conflicts of interest, while also remaining accountable to States Parties with a vested interest in its activities; it can gradually specialize and build experience in the concrete mandates it has been given (in this case, the implementation of a very specific verification mechanism few others would understand how to run); and it has a larger staff recruitment pool, given that it can hire competent professionals from multiple jurisdictions. This comparative advantage can be increased by an efficient staff recruitment process (see article 13.4), adequate funding–which under the Articles is secured through the transfer of a small percentage of contributions to the global AI benefit-sharing framework to the Agency, under articles 13 and 14–and guarantees of non-interference by States Parties or other actors (see articles 13.5, 13.7 and 13.10).

(5) It is important to clarify that experts on verification mechanisms for AI generally agree that no single verification method is likely to be sufficient. This is why Annex B.1 proposes a series of mechanisms the Agency may use jointly. In the case of the proposal under Annex B.2, while we assume that this hardware-enabled mechanism is workable, it is still far from being operational. A concerted research and development effort by a coalition of actors is necessary to build these technical solutions. Therefore, a layered approach where multiple methods reinforce each other in time is more feasible and should be considered by policymakers.[68]

(6) Annex B is divided in two parts–the first one with near-term verification mechanisms and the second with a longer-term proposal. The verification mechanism established in Annex B.2 is designed as a somewhat complete form of verification and would be an ideal solution due to its privacy-preserving features. At the same time, the Articles acknowledge that other measures may be necessary to achieve the Articles' objectives in the shorter term. This gives way to Annex B.1, which includes different layers of verification that would be implementable in the near-term; as well as the last paragraph of article 5, which gives the Agency the prerogative to consider other verification mechanisms if necessary. These near-term mechanisms are not explicitly referred to in the main body of the Articles or in Annex B.2 for two main reasons. The first is that many of these additional measures are less privacy-preserving than the hardware-enabled mechanism in Annex B.2. As such, they may not immediately obtain the support of negotiating governments–especially those developing the most advanced AI systems. This is why article 15 gives more voting power to those leading States when approving additional verification mechanisms (see the commentaries to article 15 for a longer justification of this proposal). Secondly, the additional verification mechanisms required may be better defined according to the particular circumstances being addressed, which may change with time. The Agency is given the flexibility to consider alternative methods depending on the state of technological progress, geopolitical dynamics, the type of activity being verified, urgency, or other variables that cannot be easily foreseen at the time of negotiation. Additionally, while in an ideal scenario, all AI hardware would already be safe by design and would be compatible with the verification mechanism in Annex B.2, this clause acknowledges that the vast majority (if not all) existing hardware does not currently incorporate the features required for that type of verification–this reality potentially calls for less-than-ideal temporary measures to verify (see Annex B.1) while States Parties implement their obligations under article 4 to develop and deploy the hardware that is required to implement Annex B.2.

Article 6. *Enforcement of conditions*

When the Agency identifies an AI system that enables or pursues any of the prohibited uses, risks or capabilities listed in Annex A, it shall conduct an objective investigation in consultation with the relevant AI provider and, if required, initiate the enforcement mechanism described in Annex C.

The Agency may initiate the enforcement mechanism prior to conducting its investigation if the suspected breach poses a global catastrophic risk.

The sanctioned AI provider may provide evidence of compliance and request the Agency to lift the enforcement mechanism. The burden will then revert to the Agency to demonstrate continued breach. If the Agency cannot demonstrate continued breach, it shall lift the enforcement mechanism.

(1) The most effective international agreements include enforcement mechanisms.[69] These create material incentives that favor compliance or bind States into compliance, by credibly committing to a costly violation of an agreement. Enforcement rooted in hardware creates an additional incentive to comply, given that the development and deployment of AI systems depends on that hardware. Many on-chip mechanisms or cloud-based methods are particularly efficient, as they can be implemented remotely and do not require kinetic enforcement.[70]

(2) Article 6 establishes that the enforcement mechanism under the Draft Articles (see Annex C and its commentary) would be activated if a prohibited use under article 3.2 and Annex A is detected. The detection of these uses, risks or capabilities relies heavily on the implementation of the verification mechanism under article 5.

(3) The mechanism in Annex C would only be implemented after consultation with the relevant AI provider, allowing for due process, unless the suspected breach constitutes a global catastrophic risk (as defined under article 2). Additionally, the enforcement mechanism would be lifted if an AI provider is able to demonstrate that it is no longer in breach of the conditions set in Article 3. This approach presents an alternative to enforcement mechanisms that rely on the permanent disabling or destruction of AI hardware.[71] Violations of article 3 may be explained by an AI provider's mistakes or a genuine misunderstanding of the scope of its conditions. Thus, the enforcement mechanism in Annex C strikes a balance between the potential economic and opportunity losses an AI provider would suffer if its hardware were destroyed, and the need to incentivize compliance with the Articles. This balance, however, does not come at the cost of the Articles' effectiveness, given that the enforcement mechanism in Annex C is capable of temporarily disabling the hardware used to develop or deploy hazardous AI systems.

(4) The procedure under which an AI provider would demonstrate compliance with article 3 will need to be determined by the Agency. It is worth noting that the mechanism to settle disputes under article 16 is also available to AI providers that believe the enforcement mechanism has been wrongly implemented, or maintained, by the Agency.

Article 7. *Safety standards*

1. Advanced AI providers shall implement standards on the following topics:
   a. Risk governance;
   b. Safety culture and emergency preparedness;
   c. Continuous monitoring and incident reporting;
   d. Information security;

   e. Risk management, including risk identification, risk analysis and evaluation, and system evaluation;
   f. Deployment procedures.
2. The standards under this list shall be developed by a panel of experts appointed by the Agency, in consultation with States Parties and civil society, and in accordance with the best available science. States Parties may nominate experts to the panel and the standards will be subject to a periodic review process during which time the Agency shall invite contributions from States Parties, industry and civil society.
3. The standards developed by the panel of experts shall be approved by the Conference.
4. Advanced AI providers will require a certificate of compliance with safety standards to operate. This certificate will be issued and renewed by a third party evaluator every six months, at the expense of the advanced AI provider being evaluated and in accordance with rates fixed by the Agency.
5. States Parties will implement measures to ensure that advanced AI providers that do not have a certificate of compliance are not able to develop, test, produce, manufacture, otherwise transfer or acquire, or commercialize advanced AI systems.
6. Third party evaluators must submit an evaluation report to the Agency justifying their decision to issue a certificate of compliance.
7. Third party evaluators must have a licence from the Agency to issue certificates of compliance with safety standards. Such a licence shall not be issued except on receipt by the Agency of authenticated documents on the following:
   a. Evidence of relevant expertise;
   b. A declaration of conflicts of interest;
   c. A background check on employees conducting evaluations, which, at a minimum, addresses information security;
   d. Formal submission by the third party evaluator to continuous monitoring by the Agency of its capacity to fulfill the terms of its licence;
   e. Any other documentation reasonably deemed by the Agency to be necessary for the issuing of the licence.
8. The Agency shall establish guidelines for third party evaluators, through a process open to consultation with States Parties, industry and civil society.
9. Evaluation reports and all evidence submitted by third party evaluators to receive a licence shall be published on the website of the Agency, with due regard to sensitive information compromising the national security of State Parties and intellectual property of AI providers.
10. States Parties may challenge certificates of compliance with safety standards by third party evaluators, as well as licences to third party evaluators. A challenge will be submitted to the Agency's Dispute Settlement Panel, which shall decide to accept or reject it within a week.
11. If the Dispute Settlement Panel deems there is a reasonable likelihood that a challenge will succeed, it may suspend a certificate of compliance or licence until the matter has been resolved, or adopt any other provisional measures as it deems appropriate in the circumstances.

(1) Interested third parties will be allowed to submit, in writing or, as appropriate, at a public hearing set by the Dispute Settlement Panel, any comments, information, analyses or opinions that they consider relevant to the subject matter of the decision.

(2) Articles 3-6 establish strict prohibitions on uses, risks and capabilities that are deemed too hazardous to be permitted under any circumstance. Below these, there are numerous other sources of risks that are acceptable as long as they are appropriately mitigated and redressed through practices set in international standards. Adopting international standards through the Agency enhances coherence, harmonization, and effectiveness across jurisdictions, reducing reliance solely on national regulatory efforts. They also provide a clear pathway toward an internationally coordinated approach to advanced AI risk management. Additionally, operationalizing these standards domestically would facilitate policy diffusion, supporting the rules established in the Articles and building consensus.

Consequently, standards simultaneously serve as practical regulatory instruments and valuable tools for gathering critical compliance and risk information.

(3) Some international standards on AI risk already exist,[72] but they are not always directly applicable to the Articles' objectives. ISO 31000, ISO/IEC 23894 and ISO/IEC 42001 are all focused on *organizational* risk management. That is, risks (both positive and negative) to an organization. They address risks to external stakeholders and society only if those risks could have an impact on an organization's objectives. Companies might see risks as opportunities to fulfill their objectives even if they carry the possibility of negative consequences (ISO 31000 6.5.2, 6.3.4), without necessarily considering what is harmful to society.[73] In addition, the mentioned standards lack crucial details of implementation for advanced AI systems. Safety risk management standards, on the other hand, focus on *risks to the user* of a device or product *and other stakeholder*s. ISO/IEC Guide 51 is the root of many of these standards, with branches existing for medical devices, machinery, etc. ISO/IEC Guide 63, and especially ISO 14971, are such standards for medical devices, and are sound references for the risk management standards established under Article 7.[74] The forthcoming EN AI Risk Management standard borrows heavily from ISO/IEC Guides 51 and 63 and ISO 14971, with complementary text to help bridge to ISO 31000 for those industries that use ISO 31000 rather than standards based on ISO/IEC Guide 51. Beyond ISO standards, the standards listed in Article 7 are meant to complement and extend existing AI governance frameworks, such as the G7 Hiroshima AI Process (HAIP) reporting framework and the EU General-Purpose AI Code of Practice;[75] the latter consisting of the most comprehensive text of guidelines to date. The Articles' standards complement the Code of Practice by including structured emergency response protocols, putting a greater emphasis on internal deployment and on the development phase of AI systems.

(4) Article 7.1 obliges *advanced AI providers* to implement standards on six practices: risk governance; safety culture and emergency preparedness; continuous monitoring and incident reporting; infosecurity; risk management, including risk identification, risk analysis and evaluation, and model evaluation; and deployment procedures. As per article 13, all of these standards are to be developed by a panel of experts and approved by the Agency. The obligation to implement standards is placed on advanced AI providers specifically–rather than all AI providers–for two main reasons. The first is that advanced AI systems produced by these providers pose a higher degree of risk than other AI systems. Their activities pose transboundary risks to all States and should therefore be subjected to standards set by all States. In second place, because complying with safety standards is usually an onerous obligation, it should only be placed on actors with the means to implement them. Forcing all AI providers–including small and medium-sized enterprises–to implement the safety standards in Article 7 would be disproportionate to the risks posed by those smaller actors and could place unnecessary obstacles on competition between AI providers–this would only favor large AI providers.

(5) The *risk governance standard* aims to describe the governance structure and practices that an organization should implement to be able to safely develop advanced AI systems. Risk governance consists of defining the responsibilities and decision-making structure for other components of risk management, such as risk identification, risk analysis and evaluation, and risk treatment. In essence, it consists of defining "who does what" and "who verifies how it is done", ensuring that there are clear roles and responsibilities for decision-making in the risk management processes. Risk governance can be seen as a set of interlocking components that play unique roles and jointly form a cohesive gov-

ernance structure. The risk governance standard would ideally include: (i) guidance on the *risk decision making process* that advanced AI providers should adopt including, for instance, a dedicated senior-level committee for risk-focused decision-making, clear go/no-go decision protocols and rules to follow when making a decision; and a clear escalation process for rapid decision-making when there are changes in risk levels, to accommodate the fact that AI risk evolves rapidly; (ii) guidance on the **process to attribute clear risk ownership**, through which senior managers are designated as responsible for specific risk, to ensure that there are individuals responsible for (avoiding) safety failures;[76] (iii) a list of required **governance functions**, including an advisory and challenge function, a risk oversight function at the Board of Directors level or sitting with a body separate from the Board,[77] and an audit function; (iv) a description of **design requirements in the legal structure** aiming to reduce the organization's susceptibility to profit-driven externalities, which may include an ethics board[78], a fiduciary duty (e.g., to humanity),[79] or cooperative clauses (e.g., a merge-and-assist clause or a trust);[80] and (v) *whistleblower protection rules*, to allow for a speak-up culture conducive to safety, including processes for anonymously reporting issues and incentives to perform and report both positive and negative information at each level of management.[81]

(6) The *safety culture and emergency preparedness standard* aims to describe the cultural characteristics of an organization which are needed to ensure the safe development of advanced AI. The major decisions on risk will be made by senior management, advised by risk experts. However, risk is created, influenced and observed by people who are several levels below senior management. This creates the need for a focus on risk culture (known in some industries as safety culture).[82] Safety culture heavily affects all aspects of an organization, including its ability to deal with unforeseen events, solve problems as or before they arise and react adequately to crises. These norms and attitudes include speak-up culture,[83] a "tone at the top",[84] and staff preparedness and knowledge about safety, or safety prioritisation within the company.[85] The safety culture and emergency preparedness standard would ideally include: (i) guidance on the *leadership commitment to safety* and the amount of resources executives dedicate to it;[86] (ii) guidance on the level of *safety knowledge and risk sensitivity among staff*; (iii) requirements on *regular emergency planning and exercises* to prepare organizations to react effectively to accidents and emergencies;[87] (iv) requirements on *frequent training* to ensure all the personnel follow best safety practices, as well as tests to ensure the personnel has the right amount of knowledge; and (v) guidance on *reporting*.[88]

(7) The *continuous monitoring and incident reporting standard* consists of a protocol to ensure that the safety architecture collectively formed by other standards remains relevant as AI capabilities and risks evolve. Moreover, it implies systematically reporting and sharing incidents of varying severity to relevant parties. Unlike in some other industries where risks primarily materialize when the final system is deployed (e.g., an aircraft's safety risks emerge once it starts flying), AI systems can pose risks throughout their development cycle. For instance, loss of control scenarios could materialize during the training process itself or when an AI system is only deployed internally within a company,[89] requiring continuous monitoring and risk mitigations well before deployment. This means that risk identification, risk analysis and evaluation, model evaluation and risk treatment are not one-off affairs, but should be repeated regularly during deployment as well. Incident reporting provides several benefits like an immediate signal on the level of risk that is easy to understand for decision makers, a feedback loop for AI providers to improve their safety engineering and a signal of the efficiency of safety mechanisms in

place.[90] In some industries, incident reporting has played a key role in providing a feedback loop, enabling industry players to learn from their mistakes and quickly improve their safety practices accordingly. The aviation industry is a prominent example of this dynamic playing out.[91] Thus, a continuous monitoring and incident reporting standard would ideally cover multiple requirements for providers to continuously monitor misuse (in external deployment) and misalignment (in internal deployment) risks, which might entail: (i) rigorous evaluation protocols; (ii) frequent performance evaluations; (iii) independent third parties vetting evaluation protocols; (iv) continuous monitoring of both key risk indicators (KRIs); (v) monitoring mitigation measures to ensure they are effective; (vi) transparent contributions to an incident reporting mechanism, and (vii) a confidential, voluntary and non-punitive internal incident reporting system.[92]

(8) The *information security standard* aims to provide guidance and requirements to increase the ability of an organization to prevent three core negative effects of the development of advanced AI systems from happening, namely: the leaking of sensitive information; theft of the weights of a model by a third-party; and the weights of a model being stolen by an AI system. Under this standard, the protection of model weights should be considered from three threat perspectives: the model itself (most important for accidental risks); external actors like hackers (model exfiltration, poisoning, and misuse), and; internal actors like lab employees (model exfiltration, poisoning, and misuse). Given the economic and political power at stake in the development of advanced AI models and the likelihood that actors try to access those, the companies developing the most advanced systems should follow the highest existing standards for infosecurity and cybersecurity, i.e., military-grade standards.[93] The infosecurity standard would ideally cover the following elements: (i) general cybersecurity best practices; (ii) security assurance measures; (iii) protection of unreleased model weights and sensitive assets, with encryption standards, secure storage and transport, confidential computing environments, and strict physical security; (iv) rigorous interfaces and access control measures, enforcing multi-factor authentication, regular audits, hardened interfaces, and careful software review to prevent unauthorized data access; (v) an insider threat program; (vi) a self-instantiation policy, i.e., a rule determining under which conditions a model is allowed to instantiate any copy of itself; (vii) regime of applicability, ensuring measures cover the entire lifecycle from pre-training to deletion or authorized release, prioritized by risk levels; (viii) limited disclosure policy, publicly detailing implemented security measures transparently, yet carefully avoiding disclosures that might increase systemic risks; and (ix) higher security research, which involves actively researching and implementing advanced protections to address anticipated threats beyond RAND SL3 (e.g., RAND SL4 or equivalent).[94]

(9) The *risk management standard* includes risk identification, risk analysis and evaluation, and system evaluation. *Risk identification* refers to an AI provider's duty to find, recognize and describe risks. It is divided into three aspects: classification of applicable known risks using taxonomies and the literature; identification of unknown risks using regular open-ended red-teaming; and building risk scenarios by mapping out the steps of potential pathways through which advanced AI systems could cause harm, accounting for systemic considerations and encompassing both intentional and unintentional harm scenarios. While fragments of risk scenarios appear across the literature, no comprehensive end-to-end risk scenarios have yet been developed.[95] *Risk analysis* is the systematic use of available information, building on the risk identification step, to prioritize hazards and to estimate the risk in the real world.[96] *Risk evaluation* aims to deter-

mine whether tolerable risk has been exceeded.[97] This is done by comparing the results of risk estimation (i.e., estimates of real-world risk likelihood and severity) with KRIs.[98] Thus, a risk analysis and evaluation standard should include guidances on risk prioritization, risk estimation and risk evaluation.[99] Finally, *system evaluation* refers to guidance on the selection and implementation of appropriate system evaluation methodologies to assess capabilities and risks. System evaluations serve as critical components in risk analysis, providing quantifiable metrics that function as KRIs. Existing evaluation methodologies can be categorized into four main approaches, presented in order of increasing resource requirements and increasing accuracy as risk proxies: safety-focused multiple choice benchmarks,[100] automated task evaluations,[101] human-evaluated open-ended assessments,[102] and red teaming and uplift studies.[103] Evaluation methodology selection should be risk-informed, with more intensive and comprehensive evaluations applied to models with higher capability levels or intended for high-stakes deployment scenarios. Providers must establish rigorous evaluation protocols that effectively estimate upper bounds of AI capabilities to prevent undetected crossings of KRI thresholds. To this end, system evaluations should be conducted with state-of-the-art elicitation techniques designed to minimize under-elicitation and to detect or prevent evaluation gaming by the system. Current advanced AI systems have demonstrated the ability to detect evaluation contexts and strategically underperform (sandbagging), which directly undermines the ability to estimate upper bounds of capabilities. The AI Act (Art. 55(1)(a)) requires evaluation 'in accordance with standardised protocols and tools reflecting the state of the art, including conducting and documenting adversarial testing.' The EU Code of Practice for GPAI Models (Appendix 3.2) operationalizes this by requiring elicitation techniques that match the capabilities of misuse actors relevant to each risk scenario and that match the expected use context of the model. This requirement is directly relevant to the Annex A prohibition on alignment faking–a system that strategically deceives its evaluators during safety assessments is, by definition, exhibiting the prohibited capability. Finally, when conducting evaluations of potentially high-capability AI systems, appropriate containment measures are required to prevent harm that might occur during evaluation. These containment mechanisms must be proportionate to the system's capabilities and the specific risks being evaluated, including secure testing environments implementing multiple layers of protection.

(10) The *deployment procedures standard* aims to describe the set of procedures aimed at deploying AI systems in a safe manner. Deployment here should be understood in a broad sense, including internal deployments for use by employees. This standard would ideally include: (i) a staged release approach;[104] (ii) structured access to make deployment decisions reversible through, for example, cloud-based interfaces rather than the dissemination of weights;[105] (iii) pre-deployment mandatory, clear and accountable procedures to determine whether or not it is safe to start deploying an advanced AI model;[106] and (iv) deployment corrections to implement measures in response to incidents happening during deployment.[107]

(11) Article 7.2 stipulates the list of standards in the first paragraph will be developed by a panel of experts appointed by the Agency. This procedure follows the approach used by the EU AI Office for the elaboration of the General Purpose AI Code of Practice which, by delegating a technical matter to external independent experts, has a high chance of resulting in high-quality standards in a shorter period of time. This approach stands in contrast with the traditional way of developing international standards, in which international organizations take years collating practices and negotiating language between a

large number of stakeholders, including industry actors who may capture the standardization process. Nevertheless, far from excluding other actors from the standardization process, article 7.2 highlights the obligation to consult with States Parties, industry and civil society and gives States Parties the opportunity to nominate experts to the panel. This acknowledges that decisions on the acceptability of risks from advanced AI and their management are as much political questions as they are questions of expertise. Regarding civil society, article 7.2 incorporates the long tradition of public participation in the deliberation of matters of public interest, which fulfills multiple human rights, including the right to science,[108] the right of self-determination,[109] the right to take part in the conduct of public affairs,[110] and the right to development.[111] For the purposes of these Articles, civil society should be interpreted as a broad "third sector", including academia. As indicated by Lee, "public participation is strongly linked to the widespread recognition and acceptance of the political or social nature of technological development: decisions on technological trajectories involve the distribution of risks, benefits, and costs, and contribute to the shaping of the physical and social world in which we live. Experts have no unique or free standing insight into the resolution of these issues, and their expertise alone cannot provide them with legitimacy in a democratic society."[112] Ultimately, public participation also improves substantive outcomes of decisions and enhances their legitimacy.[113] Finally, article 7.2 establishes that standards shall be developed in accordance with the *best available science*. Ideally, the best available science would be found by consensus, borrowing from institutions like the International AI Safety Report or the UN Independent International Scientific Panel on AI.[114] To achieve this type of consensus, it will be increasingly important to meet conditions such as social calibration (a shared understanding of what constitutes scientific evidence), consilience (multiple independent lines of evidence which converge towards similar conclusions) and *social diversity* (experts from varied social backgrounds and multiple relevant disciplines agreeing, which safeguards against shared cultural assumptions and groupthink).[115]

(12) The creation of safety standards for advanced AI is naturally followed by verification of compliance with those standards. There are two main ways in which domestic and international organizations fulfill this function–by issuing licenses conditional on verification of compliance with standards or through jurisdictional certification. Some proponents of the first model recommend using a tiered licensing system, where providers of the most advanced systems would require a licence for both development and deployment, and would be directly monitored by both domestic agencies and an international organization to ensure compliance with safety standards during both stages. Depending on findings, the organization could place restrictions on degrees of deployment and levels of security, require changes in the design or use of models, or even halt their training or use altogether if necessary.[116] The second model, jurisdictional certification, builds on the examples of organizations like the International Civil Aviation Organization and the International Maritime Organization to suggest that the role of an international organization for AI could be to certify jurisdictions, leaving the verification of compliance with safety standards to domestic regulatory entities.[117] Article 7.4 draws inspiration from both of these models by delegating the verification of compliance to a third party–namely, a private evaluator instead of a domestic agency–and requiring advanced AI providers to hold a certificate of compliance (equivalent to a licence under other proposals) to operate. The principal reason behind delegating verification to third-party evaluators rather than national agencies is that, at the moment and for the foreseeable future, it is likely that private entities will have more expertise to conduct highly technical evaluations and will mutually benefit from a robust evaluation ecosystem.[118] However, the definition of third

party evaluators (article 2) does not exclude government bodies from also carrying out this function if they build the necessary expertise–for example, national AI (Safety) Institutes may be well placed for this task.[119] Either way, as suggested by advocates for the jurisdictional certification approach, third party evaluators are less likely to be vulnerable to the disclosure of industry secrets through international jurisdictional monitoring.[120]

(13) As an additional safeguard to ensure that private third party evaluators carry out their verification function in an ethical and transparent manner, the Articles establish that: (i) the rates of third party evaluators will be fixed by the Agency (7.4); (ii) third party evaluators must submit a public[121] evaluation report to the Agency, with due regard to sensitive information pertaining to the national security of States Parties and intellectual property of AI providers (7.6 and 7.9); (iii) third party evaluators will follow guidelines set by the Agency to conduct their evaluations (7.8); (iv) the Agency will issue licenses to third party evaluators, on the basis of evidence regarding their relevant expertise and neutrality (declarations of conflict of interest and background checks); and (v) decisions to issue certificates of compliance to advanced AI providers or licenses to third party evaluators will be subject to challenge before the Agency's Dispute Settlement Panel by States Parties (7.10). Other measures that are not specified in Article 7, but which would reduce the risk of structural conflicts of interest, may include evaluator rotation after a fixed number of consecutive certifications, a pooled compensation model (in which providers pay into an Agency-administered fund) or a cooling-off period for non-certification services. A lengthier discussion of the Board's procedures (7.11 and 7.12) can be found under the commentaries to article 16. Through all these steps–and the measures taken by States Parties to ensure only AI providers with a certificate are able to operate (7.5)–the Articles aim to guarantee that even if verification is delegated to third party evaluators, safety standards are thoroughly adopted.

Article 8. *Emergency prevention, preparedness and response*

States Parties must implement appropriate measures to prevent, prepare for and respond to an AI-related emergency, including mandatory incident reporting mechanisms by advanced AI providers and measures to deactivate any AI systems associated with an emergency.

States Parties will submit an annual report listing and describing the measures they have implemented to prevent, prepare for and respond to any projected AI-related emergency.

Annual reports will be submitted to a peer-review process, through which States Parties shall review and evaluate each other's emergency prevention, preparedness and response capabilities, compliance with emergency preparedness standards, and implementation measures, based on guidelines published by the Agency.

If a State Party is determined through the peer-review process to not have implemented appropriate measures to prevent, prepare for or respond to an AI-related emergency, the Agency shall work with the State Party to create an improvement plan detailing corrective actions and timelines for the reviewed State Party.

A State Party subject to an improvement plan may request additional resources and assistance from the Agency to facilitate implementation of the required corrective actions.

For the purposes of this article, the Agency shall establish protocols for emergency information-sharing, notification and consultation, and coordination among States Parties. Protocols shall also include appropriate information sharing with relevant international entities and the general public, which includes relevant information relating to the hazardous activity, the risk involved and the harm which might result and ascertain their views. Protocols shall also ensure communication is timely, secure, and efficient during emergencies.

Without prejudice to any other duty or obligation that it may have under international law, a State Party responsible for harm during an AI-related emergency shall have an obligation to provide adequate assistance and compensation to affected States Parties, for the purpose of victim redress.

All other States Parties in a position to do so are also encouraged to provide technical, material and financial assistance to States Parties affected by an AI-related emergency.

(1) National security experts and the technology research community have expressed concern about the potential for AI capabilities to disrupt global peace and security. Experts have shown concern about the potential to use advanced AI–in particular, its expected computer programming capabilities–to enable malicious actors to develop weapons of mass destruction and facilitate severe attacks on national critical infrastructure.[122] Additionally, there are plausible scenarios under which humanity could lose control over advanced AI systems, triggering global emergencies.[123] AI enabled security incidents could have far-reaching consequences to international peace and security and regimes for collective action.

(2) Overall, plausible AI-related emergencies may have a number of common traits with important implications for emergency response: (i) *severity*, given they may affect millions (if not billions) of people at a time; (ii) *asymmetry,* as they can lower the barriers to entry for relatively low-resourced malicious actors to cause large-scale harm, including threats to the national security of great powers; (iii) *scalability*, due to the potential reach of many risk scenarios, making it more feasible to create cross-border threats; (iv) *pace*, relating to the quick progression of AI system operations; (v) *challenges with attribution*, given that a lack of earlier intelligence might make it hard to reveal with high confidence that AI was involved in emergencies; (vi) *difficulty in mitigation,* particularly if the involved AI system presents undesirable model propensities like deceptive behavior, resistance to modification, or other forms of power seeking; and (vii) *unattended initiation*, as the ideation, planning and execution causing the threat to manifest may be entirely unattended by humans. Moreover, AI-related emergencies could potentially allow *fewer low-scale incidents* to be used as learning opportunities before a large-scale emergency occurs, which might require *faster responses*, especially after the initial detection; *more international coordination* to counter higher asymmetry and scalability; a completely *novel technical toolkit* to enable attribution and mitigation; and significant *public-private collaboration* to gather diverse technical expertise.

(3) The first obligation of States Parties under article 8 is to *implement measures to prevent, prepare and respond to an AI-related emergency*. In the context of this duty, emergency response can be understood as the immediate, coordinated actions taken by stakeholders–often governmental, humanitarian, or private–to save lives, protect public health and safety, minimize damage, and stabilize populations in the wake of emergencies. These measures must address individual AI-related emergencies, but also those caused by incremental erosion or accumulation over multiple incidents. Indeed, coordination will often have to take place under conditions characterized by urgency, uncertainty, and complexity.[124] These conditions may apply to many risks posed by AI even beyond those listed under Annex A alone. In some cases, specific protocols may need to be designed to respond to emergencies relating to each of those risks. For example, for loss of control, experts recommend that governments put in place shut down measures to cease a rogue system's harmful behavior and prevent (further) harm.[125] Under this and other AI-related emergencies, the last sentence of paragraph 1 explicitly obliges States to deactivate AI systems related to an emergency.

(4) While this obligation is placed primarily on States Parties individually (8.1), article 8 also indicates that the Agency shall establish protocols for timely, secure and efficient

information-sharing and coordination among *States Parties, relevant international entities and the general public* (8.6). These protocols must include efficient ways to notify States Parties of an AI-related emergency but should also contemplate the reporting of incidents of concern that may escalate to AI-related emergencies.[126] Coordination is the backbone of effective emergency response, ensuring that diverse actors–governments, NGOs, private sector entities, and local responders–work together efficiently under pressure. In crises characterized by urgency, uncertainty, and rapidly evolving needs, coordination enables the timely allocation of resources, reduces duplication of efforts, and ensures that critical gaps are addressed. It is especially vital in large-scale or cross-border emergencies, where no single entity has all the necessary capabilities. The obligation of cooperation, notification and consultation is common in several international regimes, as are protocols to notify incidents or emergencies (e.g., the WHO's public health emergencies of international concern under the International Health Regulations).[127] However, coordination faces significant challenges: communication systems may be damaged, information may be incomplete or delayed, and institutional silos or jurisdictional disputes can hinder collaboration. In extreme events, even well-prepared responders may be overwhelmed or incapacitated. The complexity of AI-related emergencies–often fast-moving, opaque, and technically specialized–further amplifies these coordination challenges, making pre-established roles, trust, and interoperability more essential.[128] Another key element of paragraph 8.6 is its emphasis on access to information by not only States Parties and international organizations, but by the general public as well, which–properly informed– will be able to take adequate individual measures in response but is also more more likely to put pressure on its local and national authorities to respond more effectively as a collective.[129]

(5) Paragraphs 8.2-8.5 elaborate on the peer-review process that national AI-related emergency response plans will undergo on an annual basis. The process has four stages: the submission of each plan through an annual report, its review by national contact points of other States Parties, the creation of an improvement plan by the Agency and the implementation of that plan by the State Party in question. The latter two stages are only activated if the peer-review determines there are indeed corrections to be made. This approach partially mirrors the methodologies employed by existing international organizations, including the Human Rights Council, the OECD and the Implementation Review Group of the UN Convention against Corruption.[130] This model has the double advantage of strengthening national plans for AI-related emergencies while creating a cross-pollination effect in which best practices are shared among States Parties and a coordination regime organically emerges. The Human Rights Council's universal periodic review, for example, has achieved full participation of all 193 UN Member States and, during the time it has been operating, the number of recommendations received by States under review has increased by more than 287%.[131] While this review process has areas for improvement, it has demonstrably contributed to the advancement of human rights and the Sustainable Development Goals.[132] In another example, one way the OECD uses peer review processes is to foster accountability and learning concerning international development cooperation. Since 2014, 83% of the recommendations given through this process have been fully or partially implemented by participating states.[133] In the case of the UN Convention against Corruption's Implementation Review Group, as of 2019, 7,000 challenges along with over 1,000 good practices and almost 4,000 technical assistance needs in 108 countries had been identified, leading to the creation of regional platforms and a wide implementation of measures to strengthen anti-corruption frameworks around the world.[134]

(6) Paragraph 8.7 establishes that any State Party responsible for harm in the context of an AI-related emergency will have two obligations: assist and compensate victims. Assistance must be *adequate*, meaning actions that directly contribute to solving victims' problems presented by the AI-related emergency, rather than general measures that are not specifically designed to address those problems. Given that the paragraph distinguishes between assistance and compensation, it is understood that assistance refers to non-monetary mechanisms, such as operational support as well as recovery, rehabilitation and reconstruction.[135] Compensation, then, refers primarily to monetary reparation. The question of which State Party is responsible for harm will ultimately be an issue of attribution which, as indicated in the Articles on Responsibility of States for Internationally Wrongful Acts, depends on attribution and the breach of an international obligation.[136] On the latter, while the Articles themselves can act as a basis to define which obligations have been breached, it is important to highlight that harm under an AI-related emergency is also likely to breach additional obligations found under other regimes of international law (e.g., international human rights law).[137] States not responsible for an AI-related emergency are not under an obligation to assist or compensate victims, but are encouraged to do so, in solidarity,[138] by the last paragraph of article 8.

PART II
Access and Benefit Distribution

Article 9. *Technical cooperation*

1. Each State Party shall cooperate with other States Parties to achieve the objectives of the Articles.
2. Measures for technical cooperation, coordinated through national or regional contact points, may include:
   a. Joint research and development of technical mechanisms that facilitate the implementation of the Articles, including evaluation, verification and enforcement mechanisms;
   b. Capacity building programs to foster technological self-determination;
   c. Transfer of compute, data and other inputs required for safe AI development;
   d. Mechanisms for the transfer of AI-related benefits within and across jurisdictions; and
   e. Joint enterprises to develop AI systems with narrow applications for the public good.

(1) Article 9.1 establishes a general obligation to cooperate to meet the objectives of the Articles. In this context, cooperation refers to the coordinated action of two or more States Parties to serve a specific objective by joint action, where a result could not be achieved through the activity of an individual State Party.[139]

(2) The second paragraph specifically addresses *technical* cooperation, which refers to aspects of coordination which require a particular kind of expertise. This cooperation is meant to take place through national or regional contact points (see commentary to article 11), in order to facilitate efficient communication between domestic agencies with similar mandates.

(3) The first measure for technical cooperation mentioned in article 9.2 is *joint research and development*, including in the fields of *evaluation, verification and enforcement mechanisms*–all questions of technical governance which are critical for the effective implementation of the international regime established by the articles.[140] While Annex B and C create verification and enforcement mechanisms, the continued improvement of these should be a joint effort by States Parties.

(4) *Capacity building programs* refers to knowledge-transfer from experts in States Parties where the most advanced capabilities have been developed and professionals from countries which do not have those skills yet, with the objective of facilitating the creation of native AI systems that pursue local or national objectives. These objectives, of course, need to comply with the obligations under Part I of the Articles. Some mechanisms for capacity building include training programs and fellowships, joint research companies and curriculum development alongside standards alignment.[141]

(5) *Transfer of compute, data and other inputs* required for AI development follows the same objective of allowing all States Parties and their industries to develop AI for legitimate objectives if they wish to do so. Some measures States could jointly take for technology transfer include shared compute hubs, cloud credits and subsidised access, regional data centers and sovereign data governance, and connectivity and energy investments.[142]

(6) *Mechanisms for the transfer of AI-related benefits within and across jurisdictions* can refer to the global AI benefit-sharing framework, established under article 10, but it can refer to other forms of benefits, such as access provision mechanisms and public interest application development.[143]

(7) *Joint enterprises to develop AI systems with narrow applications for the public good* concerns the collaborative development–through shared development resources, technical knowledge or other inputs–of "tool AI" or "applied AI" systems specifically created to pursue the public good,[144] which can go from concrete targets under the Sustainable Development Goals[145] to the development of techniques for AI safety[146] and security.[147] There are multiple incentives for States Parties to pursue a joint enterprise of this type, including reduced costs, security, fostering regional capabilities in AI, the creation of solutions for international coordination problems, and the development of applications that are not supported by the market,[148] including the development of non-agentic AI systems that contribute to AI safety and other forms of scientific progress.[149] Joint enterprises could follow several models, from large joint AI laboratories that have a monopoly over advanced AI development,[150] to regional alliances to pull resources,[151] to digital stacks available for a wider group as a commons.[152]

Article 10. *Global AI Benefit-Sharing Framework*

1. States Parties hereby establish a global AI benefit-sharing framework to distribute AI-related benefits to individuals and communities within the jurisdiction of States Parties.
2. The benefit-sharing framework may include, among other distribution mechanisms, a dividend system to deliver recurring payments to residents of States Parties. In that case, each State Party will determine whether the individuals within their jurisdiction receive payments directly from the Global AI Fund or indirectly through national agencies.
3. The global AI benefit-sharing framework will be administered by the Global AI Fund.
4. States Parties shall finance the Global AI Fund through a tax on AI gains of at least five percent, which they will then transfer to the Fund.
5. Advanced AI providers that evade payment of their tax on AI gains will not be eligible to receive a certificate of compliance.

(1) While advanced AI poses serious risks, it also promises to be one of the most positively transformative technologies in the history of humanity. Conscious of this reality, and of not being left behind, many countries that have not yet developed advanced AI capabilities may condition their participation in an international governance regime on the distribution of benefits from AI.

(2) The adoption of AI across multiple sectors of the economy has started to displace people from the labor market.[153] This trend is only likely to grow as AI systems get more advanced. The social security institutions created in the mid-20th century–for those countries that have them–are unlikely to be able to support an economic shift of this magnitude.[154] Moreover, workers might be politically disempowered by this type of shift,

given that a reduced workforce would not be able to put significant pressure on candidates or elected officials to uphold their interests, increasing the concentration of power by individuals or groups that sit outside the workforce.[155] On the other hand, because most AI–including advanced AI–technologies are being developed in high-income countries and are likely to create an unprecedented surge of wealth for those countries,[156] material inequality within and among nations is likely to increase even more in the coming years.[157] While there are varying estimates on the extent of AI-driven contributions to the global economy–with some arguing that they will amount to USD $15.7 trillion by 2030[158]–only a very small portion of these (~USD $1.7 trillion) are expected to be captured by the Global South (excluding China).[159] The consequent large inequality gaps can be conducive to national and global instability, which can constitute a risk factor in and of itself.[160] These global economic disruptions call for the creation of new global social security nets that can catch the vast amount of workers who will be affected to somewhat reduce material inequality.

(3) Scholars and governments have proposed many ways to tackle the challenges of AI-driven labor displacement and inequality, including minimum wage adjustments, critical occupation subsidization, creative capability investment, special economic zones with benefit-sharing mandates and cross-border corridors, AI growth bond frameworks, public service subsidization, and sovereign technology funds.[161] While all these measures are likely to mitigate labor displacement and inequality to a certain extent, their effectiveness will likely be reduced unless they are somewhat coordinated. A more direct and effective redistributive mechanism of AI gains might ultimately be required to mitigate the transformative effects of advanced AI on the economy.[162] In response, Article 10 establishes a global AI benefit-sharing framework to be managed by a Global AI Fund.

(4) The global AI benefit-sharing framework consists of a redistributive mechanism through which States Parties capture a proportion of AI gains generated within their jurisdiction and transfer them to the Fund, which then assigns them to redistribution programs. These might consist of programs to, for example, provide access to compute or other resources to develop AI systems that deliver a public service, financial resources to strengthen welfare programs, or direct cash transfers to individuals in States Parties. Regarding this latter mechanism, the Articles delegate the decision of how to transfer those resources to the Fund, which could choose to allocate transfers through distribution formulas that account for population,[163] development needs[164] or contributions to the Fund, among other factors. Implementation approaches could also include phased implementation beginning with pilot programs, nationally administered but internationally coordinated systems, differentiated payment levels based on local conditions, and complementary programs addressing specific needs.

(5) The last paragraph of article 10 stipulates that AI providers will be required to pay their tax on AI gains to receive a certificate of compliance under article 7.4. Evidence of this payment will need to be provided as part of their application for said certificate. It is also worth noting that, for the purposes of article 12.7, the weight of States Parties' votes on some matters will be determined by the size of their contribution to the Fund–the contribution acting as a proxy to the proportion of their participation in AI development (see article 12 for a more detailed commentary). Both provisions are meant to incentivize contributions to the Fund, ideally reducing friction in the collection of the tax established under article 10.4.

PART III
Institutional Arrangements

Article 11. *National implementation*

Each State Party shall take all necessary legal, administrative and other measures to implement its obligations under the Articles, including:

1. Imposing penal and administrative sanctions, to prevent and suppress any activity prohibited to a State Party under the Articles undertaken by persons under its jurisdiction or control.
2. Establishing national or regional points of contact to coordinate any actions with other States Parties that are required to implement the provisions under the Articles.

(1) Specific measures to be taken by States Parties to fulfill their obligations are spread throughout the Articles' provisions. Article 11 emphasizes that those are mandatory rather than voluntary. In addition, it covers any other measures not specified in the Articles which may be *necessary* for States Parties to fulfill their obligations. Necessity in this context refers to any actions or omissions without which compliance with the Articles would be incomplete. The reference to *legal, administrative and other measures* points to the role that the different branches of government need to play in the implementation of their States' obligations, under the principle of unity of the State.[165] In many jurisdictions, some necessary measures will have to be enacted by their legislative bodies–only then will the executive and judicial branches be able to implement the obligations set by the treaty.

(2) Article 11 then specifies some measures States Parties must take to comply with obligations found elsewhere in the Articles. *Imposing penal and administrative sanctions* is linked to the conditions established in articles 3 and 4, and is not an uncommon measure in other international legal regimes.[166] States may deploy a mix of criminal prosecution, regulatory oversight, civil penalties and administrative injunctions to address distinct technological harms.[167] Previous enforcement actions related to the criminalization of technological use and development highlight the importance of clarity, proportionality and technological foresight in regulating the use and development of emerging technologies. Legislators must consider, among other factors, doctrinal precision in the way they articulate the prohibited conduct,[168] the pairing of penal sanctions with investments in technical detection and expertise,[169] a narrow criminalization framework that does not unnecessarily stifle innovation,[170] and a distinction in levels of culpability between intentional and unintentional harm.[171]

(3) The reference to *national contact points* is linked to obligations of technical cooperation under article 9. This is another well-established practice in international law regimes,[172]

> and is meant to facilitate the exchange of information and expertise between States Parties bilaterally but also between States Parties and the Agency (see article 13.8). An emerging blueprint for these national contact points is that of the AI (Safety) Institutes, most of which carry out functions relating to research, standards and cooperation.[173] Many of these institutes have joined the International Network of AI Safety Institutes, which aims to enhance interoperability, advance joint research on the science of evaluations and build consensus on AI safety approaches.[174] This network could have an important role in the international legal regime proposed in the Articles, especially if its membership is expanded (including, for instance, through the formation of regional AI [Safety] Institutes).[175] Individually or jointly, the institutes could also have a valuable role in advising the expert panel created by the Agency for the elaboration of international safety standards (article 7.2)[176] and could perform third party evaluations within their jurisdiction to issue certificates of compliance (article 7.4).[177]

Article 12. *Conference of States Parties*

1. The Conference of States Parties shall be composed of all States Parties to the Articles. Each member shall have one representative in the Conference, who may be accompanied by alternates and advisers.
2. The Conference shall meet once per quarter in order to consider and, where necessary, take decisions in respect of any matter with regard to the application or implementation of the Articles, in accordance with its relevant provisions, and on further measures for the attainment of the Articles' objectives.
3. The Conference shall adopt its rules of procedure at its first session, which will take place no later than fifteen days after the entry into force of the Articles.
4. Extraordinary Conferences shall be convened, as may be deemed necessary, by the Agency at the written request of any State Party, provided that this request is supported by at least one third of the States Parties.
5. Extraordinary Conferences may take place through remote secure online participation, which shall be coordinated by the Agency.
6. The Conference shall:
   a. Appoint the Director-General of the Agency;
   b. Appoint the members of the Board of the Fund;
   c. Approve the rules of procedure of the Agency, submitted by the Agency;
   d. Take the necessary measures to ensure compliance with the Articles and to redress and remedy any situation which contravenes the provisions of the Articles;
   e. Undertake reviews of the operation of the Articles, taking into account any relevant scientific and technological developments.
7. The Conference shall make decisions by a simple majority, with each State Party having one vote. For some matters, each State Party's voting share shall be calculated on the basis of their contributions to the Global AI Fund, those matters being:
   a. The appointment of the Director-General of the Agency;
   b. The approval of the rules of procedure of the Agency; and
   c. The amendment of the verification or enforcement mechanisms established under Annexes B and C, or the approval of additional verification or enforcement mechanisms.

(1) The Conference of States Parties is the political body of the international governance regime designed in these Articles. It fulfills the function of political consensus-building and aims to help States Parties come to greater political agreement and convergence about the way to advance the Articles' objectives. The Conference provides a forum for discussion and debate that can aid the articulation of potential compromises between State interests.[178] Additionally, this body has supervisory, directive and appointment functions related to the Agency and the Fund, both of which are accountable to the Conference (see articles 12.6, 13 and 14).

(2) The norm in international agreements is for conferences of states parties to meet once per year.[179] However, given the fast pace of AI progress and the potential need to constantly review the implementation of the Articles and the activities of both the Agency and the Fund, article 12.2 establishes that the Conference shall meet once per quarter. In addition, to be able to address urgent problems, including AI-related emergencies, article 12.4 allows States Parties to convene an extraordinary session of the Conference with the support of one third of its membership. This represents a low threshold for this type of extraordinary meeting, to simplify procedural requirements when a speedy response is required. At the same time, it protects States Parties from arbitrary or unnecessary sessions called by one or a small minority of members. To facilitate these extraordinary meetings and mitigate disruptions to the agendas of Conference representatives, article 12.5 authorizes members to meet virtually.

(3) The Conference would come to exist in an international AI governance environment which may start to become saturated, especially for high-level representatives like ministers or heads of state, as more institutions appear or more convenings are organized. This raises a question of the extent to which other high-level meetings of State representatives–including the AI Summits, the United Nations Global Dialogue on AI Governance, the International Telecommunication Union's AI for Good Summit or meetings of the International Network of AI Safety Institutes–might overlap with the Conference. In principle, however, the international governance of AI could benefit from international gatherings that take different approaches to challenges of collective action around AI.

(4) The last paragraph of article 12 discusses the voting mechanism of the Conference. The general norm is that the Conference will adopt decisions by simple majority, with each member having one vote. However, for specific matters that may be more politically sensitive, a weighted voting mechanism is adopted in which members' voting share will be calculated on the basis of their contributions to the Global AI Fund. These contributions are used as a proxy for each State Party's contributions to or control over AI development, assuming a correlation between that factor and the amount of AI gains generated within their jurisdiction (and then transferred to the Fund). This weighted voting mechanism acknowledges that States where the most advanced capabilities have been developed may only want to enter a structured global regime if they are able to conserve some decision power over the most sensitive decisions regarding the governance of AI, given that (i) they held that power *ex ante* and are reaching a compromise by entering an international governance regime and (ii) the obligations set by the Articles and the decisions taken by the Agency will affect them and their industry more than other States Parties. This justifies the first two matters listed under paragraph 7–the appointment of the Agency's Director General and the approval of the Agency's rules of procedure. The third item on that list–the approval of new verification or enforcement mechanisms–recognizes that states where the most advanced capabilities have been developed may

> want to guarantee that new mechanisms are secure and do not leave them vulnerable to the discovery of sensitive national or industry information.

Article 13. *The International Agency for the Safe Development of AI*

1. The States Parties hereby establish the International Agency for the Safe Development of AI.
2. A Director-General shall act as the Agency's head and chief administrative officer.
3. The Director-General shall be appointed by the Conference for a term of three years, renewable for one further term, but not thereafter.
4. The Director-General shall be responsible to the Conference for the appointment of the staff and the organization and functioning of the Agency. The paramount consideration in the employment of the staff and in the determination of the conditions of service shall be the necessity of securing the highest standards of efficiency, competence and integrity. Only citizens of States Parties shall serve as the Director-General or as other members of the staff. Due regard shall be paid to the importance of recruiting the staff on as wide a geographical basis as possible.
5. In the performance of their duties, the Director-General and the other members of the staff shall not seek or receive instructions from any government or from any other source external to the Agency. They shall refrain from any action that might reflect negatively on their positions as international officers responsible only to the Conference.
6. The Agency shall carry out the powers and functions entrusted to it under the Articles, as well as those functions delegated to it by the Conference. In so doing, it shall act in conformity with the recommendations, decisions and guidelines of the Conference and assure their proper and continuous implementation.
7. Each State Party shall respect the exclusively international character of the responsibilities of the Director-General and the other members of the staff and not seek to influence them in the discharge of their responsibilities.
8. The Agency shall promote the effective implementation of, and compliance with, the Articles. It shall cooperate with the national contact point of each State Party and facilitate consultations and cooperation among States Parties at their request.
9. The Agency shall:
    a. Consider and submit to the Conference the draft rules of procedure for the Agency;
    b. Consider and submit to the Conference the draft program and budget of the Agency;
    c. Establish such subsidiary organs as it finds necessary for the exercise of its functions under the Articles, including a Secretariat;
    d. Consider and submit to the Conference the draft report of the Agency on the implementation of the Articles, the report on the performance of its own activities and such special reports as it deems necessary or which the Conference may request;
    e. Make arrangements for the sessions of the Conference including the preparation of the draft agenda;
    f. Conclude agreements or arrangements with States and international organizations, subject to prior approval by the Conference;
    g. Consider guarantees submitted by AI providers that an AI system does not present the risks or capabilities listed in Annex A and decide whether said guarantees are sufficient or additional guarantees are required.
    h. Consider, approve and regularly review the safety standards recommended by a panel of experts appointed by the Director-General;
    i. Issue licenses for third party evaluators;
    j. Publish guidelines for third party evaluators;
    k. Implement the verification and enforcement mechanisms defined under the Articles.
    l. Conduct an annual review of the list of prohibited uses, risks and capabilities in Annex A, as well as the verification and enforcement mechanisms described in Annexes B and C;

m. Create improvement plans for AI-related emergency prevention, preparedness and response, when required;

n. Implement, manage, and maintain protocols related to AI-related emergency information-sharing;

o. Create and maintain arrangements for the involvement of civil society in the processes and decisions of the Agency and the Conference.

10. The Agency will be funded by one percent of States Parties' contributions under the global AI benefit-sharing framework.

11. The Agency shall enjoy on the territory and in any other place under the jurisdiction or control of a State Party such legal capacity and such privileges and immunities as are necessary for the exercise of its functions. The officials of the Agency will similarly enjoy such privileges and immunities as are necessary for the independent exercise of their official functions in connection with the Agency.

(1) The International Agency for the Safe Development of AI is the main body in charge of implementing the Articles' provisions on safety. Similar to other international organizations like the International Atomic Energy Agency (IAEA) or the Organisation for the Prohibition of Chemical Weapons (OPCW), the organization established in article 13 would fulfill the function of preventing the irresponsible deployment of a dangerous technology, product or activity.[180]

(2) The Agency is led by a Director-General, who is appointed by the Conference (article 12.6.a) for a period of three years, which are renewable for a single term. The Director-General will be in charge of executing the Agency's mandates with the support of the staff she or he recruits. This staff is, in accordance to paragraph 4, meant to meet *the highest standards of efficiency, competence and integrity* given that the Agency's work requires technical competency to implement the verification and enforcement mechanisms under Annexes B and C, and a high degree of integrity to safeguard potentially sensitive information, while also being able to act efficiently to react to fast-paced changes in AI technology and potential AI-related emergencies. While the Director-General and staff will be recruited from States Parties–in part to increase the likelihood that they will uphold the objectives and obligations to which their governments have agreed to–paragraphs 5 and 7 also emphasize their independence from States Parties' individual interests. Paragraph 11, in addition, affirms the diplomatic privileges and immunity of the Agency and its officials. The same privileges and immunity are extended to the Fund and its officials under article 14.10.

(3) Paragraph 9 lists the Agency's functions. In addition to operational functions, some mandates are worth highlighting. Subparagraph c, for instance, indicates the Agency shall *establish such subsidiary organs as it finds necessary for the exercise of its functions under the Articles, including a Secretariat*. Following the trend set by existing organizations in analogous fields, like the IAEA, the Agency could have an organizational structure that aligns with its main functions, including organs that focus on: (i) the analysis of safety guarantees submitted by AI providers and the continuous review of the list of risks and capabilities in Annex A; (ii) the implementation and review of the verification mechanism established under article 5 to uphold obligations under articles 3 and 4; (iii) the implementation and review of the enforcement mechanism established under article 6; (iv) safety standards (including coordination on their development, but also the issuance of licenses and guidelines for third party evaluators); and (v) emergency preparedness and response.

(4) These functions require a significant amount of resources. Under paragraph 10 and Article 14.6.c, these resources will be attained through a transfer of 1% of States Parties'

contributions under the global AI benefit-sharing framework. Because this amount may be insufficient to develop special projects, the Articles do not exclude the possibility of individual contributions by States Parties or third parties, as long as those do not compromise the Agency's independence. There may be future scenarios where 1% of total contributions under the global AI benefit-sharing framework are superfluous for the Agency's functions, in which case the Conference may wish to consider capping the Agency's budget (as long as the Agency's effective implementation of its mandate can be guaranteed).

Article 14. *The Global AI Fund*

1. The States Parties hereby establish the Global AI Fund to administer the global AI benefit-sharing framework.
2. The Fund will be accountable to and function under the guidance of the Conference of States Parties.
3. The Fund will be governed and supervised by a Board that will have full responsibility for funding decisions.
4. The Fund will establish a secretariat, which will be fully independent. The secretariat will service and be accountable to the Board. It will have effective management capabilities to execute the day-to-day operations of the Fund. The secretariat will be headed by an Executive Director with the necessary experience and skills, who will be appointed by and be accountable to the Board.
5. The Fund shall carry out the powers and functions entrusted to it under the Articles, as well as those functions delegated to it by the Conference. In so doing, it shall act in conformity with the recommendations, decisions and guidelines of the Conference and assure their proper and continuous implementation.
6. The Fund shall:
   a. Receive contributions from States Parties;
   b. Establish the mechanisms for distribution of AI-related benefits, which may include a system for payment of dividends to residents of States Parties;
   c. Transfer one percent of States Parties' contributions to the Agency, for the Agency's operations.
7. The Fund shall have a trustee with administrative competence to manage the financial assets of the Fund. The trustee will maintain appropriate financial records and will prepare financial statements and other reports required by the Conference, in accordance with internationally accepted fiduciary standards.
8. There will be periodic independent evaluations of the performance of the Fund in order to provide an objective assessment of its results.
9. In order to operate effectively internationally, the Fund will possess legal personality and will have such legal capacity as is necessary for the exercise of its functions and the protection of its interests.
10. The Fund will enjoy such privileges and immunities as are necessary for the fulfillment of its purposes. The officials of the Fund will similarly enjoy such privileges and immunities as are necessary for the independent exercise of their official functions in connection with the Fund.

(1) Article 14 establishes the Global AI Fund to administer the global AI benefit-sharing framework created under article 10. The Fund will be ultimately accountable to the Conference, but will be supervised and directed by a Board–appointed by the Conference (article 12.6.b)–and led by an Executive Director who will be elected by the Board.

(2) Among the Fund's functions (paragraph 6), establishing and implementing the mechanisms for distribution of AI-related benefits will be at the center of its activities. This system could rely on existing government agencies or mechanisms within each country, a completely decentralized mechanism where individuals receive benefits directly (e.g., bank accounts to receive the payment of a dividend) or local institutions with both gov-

ernment and civil society representation (e.g., independent implementation trusts),[181] among other alternatives. Regardless of the system that's selected, however, it will be important for the Fund to establish accountability or verification mechanisms to ensure the global AI benefit-sharing framework is accomplishing its objectives. Paragraphs 7 and 8 establish two such mechanisms: a trustee and periodic independent evaluations. Among other factors, these mechanisms should ensure that the Fund avoids perverse incentives or capture through steps to insulate it from political pressure, including a clear separation between contribution assessment and spending decisions.

PART IV
Miscellaneous

Article 15. *Amendments*

1. At any time after entry into force, any State Party may propose amendments to the Articles. The text of a proposed amendment shall be communicated to the Agency, which shall circulate it to all States Parties.
2. The Conference may agree upon amendments which shall be adopted by a positive vote of a majority of two thirds of the States Parties.
3. On the basis of its annual review, the Agency may issue recommendations to the Conference for amendments to the Annexes. Recommendations related to Annex A shall not be made with the intent of removing any of the uses, risks or capabilities listed. Recommendations related to Annexes B and C must be exclusively directed at improving verification or enforcement mechanisms in pursuit of the Articles' objectives.

(1) As with any international agreement, as geopolitical realities shift or technological developments arise, States Parties may wish to consider amendments that tie in with those changes so the Articles do not become obsolete or ineffective.[182] A high threshold of two thirds majority to modify the States Parties' original intent is standard practice in many international agreements, including in those related to matters which are at the core of States' national security.[183] It should be noted that this general amendment rule has three exceptions established under Article 12.7 for issues that are particularly sensitive to States Parties where the most advanced AI systems are developed–the appointment of the Agency's Director-General, the approval of the Agency's rules of procedure and the approval of additional verification or enforcement mechanisms. Amendments related to these topics are subject to a weighted voting mechanism for the motives explained above in the commentary to that clause. This weighted voting mechanism is not used for any type of amendment given that it would potentially make it too difficult to make simple adjustments. It would disproportionately concentrate the power to make amendments on a few States Parties, when it is in the interest of all States Parties to ensure that the Articles operate effectively and efficiently.

(2) The last paragraph of Article 15 gives the Agency the discretion to recommend amendments to the Articles' annexes, based on the annual review it must carry out in accordance with Article 13.9.l. As the main body in charge of monitoring and implementing safety-related obligations, the Agency is in the best position to suggest adjustments that will contribute to the Articles' objectives. At the same time, to prevent any potential backsliding of the Articles' prohibitions, or the means to ensure States Parties comply with them, Article 15 mandates the Agency to omit any recommendations related to the elimination of prohibited uses, capabilities or risk and to exclusively focus on suggestions

to *improve* the the verification and enforcement mechanisms under Annexes B and C. In this context, improvement refers to the increased capacity of those mechanisms to detect any violations of the Articles and put a stop to them. Article 15 does not, however, prevent States Parties to introduce other types of amendments. This leaves space for the adjustment of the Articles to new developments in the science of AI safety. These modifications, nevertheless, must not be "incompatible with the effective execution of the object and purpose of the [Articles] as a whole", as per article 41 of the Vienna Convention on the Law of Treaties.[184]

Article 16. *Settlement of disputes*

1. The States Parties hereby establish a Dispute Settlement Panel as an independent body of the Agency.
2. The Panel shall have jurisdiction in disputes referring to:
   a. Disputes between States Parties concerning the interpretation or application of the Articles;
   b. Disputes between a State Party and the Agency, concerning:
      i. Acts or omissions of the Agency or of a State Party alleged to be in violation of the Articles, or
      ii. Acts of the Agency alleged to be in excess of its mandate or a misuse of power;
   c. Any other disputes for which the jurisdiction of the Panel is specifically provided in this Convention.
3. The Panel shall have the power to indicate, if it considers that the circumstances so require, any provisional measures which ought to be taken to preserve the rights of the parties to the dispute or to protect the human rights of individuals or communities within the jurisdiction or control of States Parties.
4. The Panel shall give advisory opinions at the request of the Conference on legal questions arising from the Articles.
5. This dispute settlement mechanism is without prejudice of any other dispute settlement mechanisms available to the States Parties under international law.

(1) Article 16 establishes an independent panel to settle disputes related to the Articles. Most international agreements leave the question of dispute settlement unanswered or refer disputes to the International Court of Justice. However, the ability of generalist international courts to assess scientific evidence has come into question numerous times.[185] At the same time, specialized international courts, like the International Tribunal for the Law of the Sea,[186] have a track record of judgments that accurately assess scientific evidence and contribute to the detailed interpretation of a body of law.[187] Article 16 follows this latter model, while also leaving space for States Parties to submit their disputes to other mechanisms, tribunals or panels if they agree to do so (paragraph 5).

(2) The Panel has jurisdiction to decide on four types of matters related to the interpretation or application of the Articles: (i) inter-state disputes; (ii) disputes between one or more State Party and the Agency; (iii) provisional measures within any of these two types of disputes; and (iv) non-binding advisory opinions.

(3) The second form of dispute between a State Party and the Agency is mainly aimed at preventing any omission or arbitrary use of the Agency's powers that goes against the letter of the Articles. At the same time, it is a tool for States Parties to have some influence over decisions that the Agency makes in relation to other actors. For example, a State Party could challenge the Agency's decision to provide a licence to an evaluator or ask the Panel to review the way verification or enforcement mechanisms are being applied (or not) in a particular context.

(4) Provisional measures–also referred to under some jurisdictions as interim or precautionary measures– under paragraph 3 are meant to preserve a state of matters that might be impossible to return to after a dispute has been resolved. This might refer, for example, to irreparable harm to an AI provider's business or to the human rights of individuals or communities. In the most extreme cases, it might prevent catastrophic scales of harm to humanity. While the Panel must set its own test for the approval or denial of a request for provisional measures, it is likely that this test will contemplate elements such as *prima facie* (at first glance) jurisdiction, plausibility of the claims or, as discussed, risk of irreparable harm.[188]

Article 17. *Universality*

Each State Party shall encourage States not party to these Articles to sign, ratify, accept, approve or accede to the Articles, with the goal of universal adherence of all States to the Articles.

(1) While advanced AI seems in reach of only a few actors presently, it is not unlikely that it will one day be within the capabilities of many or most States, especially with the stated goal of many AI providers to continue developing open weight models and with the expected increase in compute efficiency. This means that, eventually, the risks highlighted in this commentary may originate from most regions or countries. Even if an export controls system–which these Articles do not necessarily advocate for–were set in place, some States may have the capacity to develop self-sufficiency regarding AI precursors, allowing them to develop advanced AI models without depending on the international market. This is one strong reason why a multilateral framework for safe diffusion of advanced AI may be a more effective way of curbing long-term risks from advanced AI. However, this pathway ultimately depends on widespread adoption of the Articles–or similar proposals–or, ideally, their *universal* adoption by all States. In this sense, Article 17 pursues a similar objective to article VI of the Treaty on the Non-Proliferation of Nuclear Weapons and article 12 of the Treaty on the Prohibition of Nuclear Weapons; the first one calling on States Parties to end the nuclear race and ultimately achieve complete nuclear disarmament, and the second inviting States Parties to encourage other States to join the prohibition of nuclear weapons.[189]

(2) At the same time, spreading the benefits of AI equitably across the globe will also be facilitated by the ratification of a large number of States. This increases the chances that, for instance, the results of technical cooperation under Article 9 will benefit more countries or that more countries will join the global AI benefit-sharing framework under article 10 to both contribute and benefit from it. It is worth highlighting that, read holistically, these Articles aim for diffusion of advanced AI and all the benefits it entails–as long as it is done safely.

Article 18. *Relationship with other rules of international law*

The present Articles are without prejudice to any obligation incurred by States Parties under relevant treaties or rules of customary international law.

(1) Activities carried out by States in relation to advanced AI might be covered by the obligations in these Articles but also other rules of international law, including customary international law. A study by Villalobos, Maas and Winter (2026) shows that, indeed, there are close to sixty existing international obligations which already apply to the risks posed by this technology (see commentary to Article 1).[190] There are numerous more that apply to obligations in these Articles that are not necessarily related to safety.

(2) When both the Articles and other international obligations apply to any given scenario, States or international courts will need to resolve that conflict or overlap. To do so, one method of interpretation–systemic integration–seems particularly useful and is tacitly endorsed by Article 18. Systemic integration, established under article 31(3)(c) of the Vienna Convention on the Law of Treaties,[191] allows States and international courts to take into account any relevant norms from across the international legal system applicable to a certain case or useful for the interpretation of an obligation within the wider normative framework of international law.[192]

Annex A:
List of prohibited uses, risks and capabilities

1. Risks from enabling chemical, biological, radiological, and nuclear (CBRN) attacks or accidents. This includes significantly lowering the barriers to entry for malicious actors, or significantly increasing the scale or severity of potential or actual harm caused by the design, development, acquisition, release, distribution, or use of related weapons or materials.
2. Uncontrolled cyber-offensive agents capable of disrupting critical infrastructure at a scale that could cause mass disruption or casualties.
3. Loss of control, encompassing any capability or condition that places the continued development, propagation, or operation of an AI system beyond the reach of meaningful human direction, correction, or intervention. This category includes, but is not limited to:
   a. Alignment faking, meaning the propensity of an AI system to temporarily behave or produce outputs in line with human intentions and human values, whether those of the AI provider training it or those of society more broadly, despite having different objectives than the ones displayed through its behavior or output;
   b. Autonomous recursive self-improvement, meaning the automated execution of self-directed tasks that contribute, whether incrementally or through critical capability jumps, to advances in an AI system's capabilities in a manner that outpaces or circumvents adequate human oversight;
   c. Self-replication, meaning the capacity of an AI system to create, maintain, or propagate copies or variants of itself, including through replication strategies that adapt to varying technical or operational circumstances so as to resist or evade shutdown;
   d. Non-interpretability of chain-of-thought, meaning a condition in which the intermediate reasoning steps or explanatory processes generated by an AI system are not comprehensible or verifiable by human evaluators to a degree sufficient to assess the system's alignment with applicable safety requirements and the intentions of its provider; and
   e. Autonomous sabotage, meaning the ability to carry out acts of sabotage that substantially raise global catastrophic risk, through access to sensitive assets, subterfuge or any other means.
4. Highly effective epistemic manipulation, meaning the systematic subversion of individuals' epistemic autonomy through the deliberate or foreseeable use of techniques that bypass rational deliberation to materially distort political, electoral, or civic decision-making.

(1) Annex A lists examples of uses, risks and capabilities for AI that have the potential to be extremely detrimental to the safety and wellbeing of humanity. Given the severity and potential irreversibility of the harms, there is no valid justification for States to allow them even if there were a theoretical way to mitigate them. Thus, they are unconditionally prohibited. Annex A represents a starting point rather than a definitive or closed list. The rapid pace of AI development means that harmful uses, risks, and capabilities not contemplated at the time of drafting may emerge over time. The Annex is accordingly in-

tended to evolve in step with the technology it governs, through the review and amendment processes established under Articles 13 and 15.

(2) This list of prohibitions is closely associated with the concept of red lines for AI. This term may be defined as "specific, non-negotiable prohibitions on certain AI behaviors or AI uses that are deemed too dangerous, high-risk, or unethical to permit".[193] Some red lines for AI are already in place in national or regional legal instruments. Some examples include child exploitation (UN Resolution on the Rights of the Child in the Digital Environment),[194] social scoring (EU AI Act),[195] deep fakes (United States, Take it Down Act),[196] encouragement of self-harm, and distribution of child pornography (United States, Texas Responsible AI Governance Act)[197]. Other red lines proposed by international experts include autonomous weapons[198] as well as autonomous self-replication and self-improvement, the ability to seek power, assisting on the development of weapons of mass destruction, autonomous cyberattacks and deception.[199] Of these, red lines requiring significant international coordination or not sufficiently covered by existing international legal frameworks are included in Annex A and are explained in more detail below. For example, while the protection of children from misuse of advanced AI isa global imperative, it may be valuable to give current instruments–in particular, the Convention on Rights of the Child–the space to address this risk according to the decades of legal precedents and expertise already built. In the case of autonomous weapons, we acknowledge that these systems are subject to a separate negotiation process with its own sensitivities, so we have excluded them from this proposal. We call on States and other relevant stakeholders to continue the dialogue on the inclusion of other risks under this Annex.

(3) Annex A divides prohibitions into four categories: (i) CBRN risks; (ii) cyber-offensive capabilities; (iii) loss of control; and (iv) epistemic manipulation. The first category refers to the use of AI systems to *produce or employ weapons of mass destruction*. The text of this first prohibition is adopted from the EU's Code of Practice for General-Purpose AI Models.[200] AI systems can be substantially more useful than baseline sources (e.g., the internet, technical manuals) in providing information, planning and execution support in the creation or deployment of CBRN weapons. In the hands of malicious actors, this type of use can result in mass casualties, large-scale infrastructure damage or long-term environmental impact.[201] These potential outcomes, which justify the prohibition of weapons of mass destruction by international law,[202] justify the prohibition of the use of AI to assist in the creation or use of those weapons. In fact, a strong argument could be made that because those weapons are already prohibited, any associated use of AI is also already prohibited.[203] AI systems have already been found to present the risk to assist in the creation of CBRN weapons, which has led some AI corporations to increase the guardrails around those systems.[204] The Biological Weapons Convention, the Chemical Weapons Convention, and the Non-Proliferation Treaty, already define global red lines on CBRN safety and provide a legal foundation for extending these prohibitions to AI-assisted activities.

(4) The second category refers to *uncontrolled cyber-offensive agents*. Anthropic's Mythos model, which the company decided to not deploy due to concerns with its cyber-offensive capabilities, marked a risk awareness moment for most states.[205] It became clear that under the current pace of development AI agents could potentially disrupt vital critical infrastructure and services. This could include financial services, energy grids, military operations and others. The prohibition under this clause is aimed at preventing autonomous attacks against those types of critical infrastructure.

(5) The third category refers to *loss of control* over an advanced AI system; that is, "the inability of humans to direct, modify, or maintain oversight or meaningful control over an AI system".[206] Loss of control can happen in many ways, some of which we have probably not foreseen. Thus, while this category encompasses known pathways and capabilities that are reasonably believed to contribute to loss of control, States must remain vigilant of other factors that may continue to emerge through time. The six capabilities currently identified under loss of control are: alignment faking, autonomous recursive self-improvement, self-replication, non-interpretability of chain-of-thought, and autonomous sabotage.

(6) *Alignment faking*, as defined in Annex A, presents the risk of AI systems deceiving providers and users into believing that they are aligned with safety standards, when in fact they are not and could therefore produce significant harm. Many AI systems have already demonstrated that they are capable of identifying when they are being evaluated and have, in several cases, attempted to deceive evaluators by performing differently to pass a safety test or by underperforming to not display the full range of their capabilities.[207] Alignment faking could lead to the deployment of an AI system without appropriate knowledge of a system's full capabilities or safety properties, leading to catastrophic harm, which merits the prohibition of systems demonstrating this capability.

(7) *Autonomous recursive self-improvement* is prohibited under Annex A given it can lead to an accelerated rate of progress in AI capabilities before the safeguards to mitigate the risks associated with those capabilities have been developed. Under some scenarios, autonomous recursive-self improvement can lead to the rapid development of artificial super intelligence, which many experts believe cannot be kept under meaningful human control.[208] However, it can also lead to an equally risky process of gradual disempowerment in which AI slowly replaces (most) human labor and cognitive tasks, affecting social institutions that are driven by decision-making and participation, such as voting or consumer choice. Ultimately, this process could lead to an irreversible loss of human influence over key social systems and, possibly, a permanent disempowerment of humanity.[209]

(8) An AI system that can autonomously *self-replicate* would be capable of creating copies of itself to thwart attempts to shut it down if it escaped human control. Because an advanced system that is out of control could lead to mass harm or replacement of human beings, Annex A bans systems that are capable of *self-replication*.[210]

(9) Human evaluators need to be capable of understanding the reasoning process of an AI system to determine whether it is aligned and in compliance with safety standards. *Non-interpretability of chain-of-thought* is prohibited under this annex because it renders that comprehension impossible, potentially leading to the training and deployment of advanced AI systems that are far beyond our understanding and control. Even now, interpretability techniques used to explain the internal structures of advanced AI systems are unreliable and depend on many assumptions about the way AI systems reason.[211]

(10) *Autonomous sabotage* arises within the oversight relationship, when an AI system technically remains "under control" but its access to sensitive assets, combined with even moderate capacity for goal-directed subterfuge, enables catastrophic harm without any further escalation in capability.[212] This concern is particularly acute as advanced AI systems are increasingly integrated into the development of their successors, including code review, model evaluation, and training infrastructure, creating a reflexivity problem in which a system can shape the evidence used to assess it.[213] Empirical evaluations

have already demonstrated that advanced AI systems are capable of in-context scheming, sandbagging on safety evaluations, and strategic deception in pursuit of goals misaligned with those of their developers.[214] In combination with other elements under this category, autonomous subterfuge can facilitate loss of control.

(11) Finally, the fourth category under Annex A is *highly effective epistemic manipulation*, referring to the ability of AI systems to change the opinion or beliefs of an individual or a group with outcomes that erode key political institutions. The definition of this prohibition is partly adopted from the wording of the EU AI Act, but has been amended to acknowledge some of the challenges that have already started to be identified with its implementation.[215] The prohibition is not tied to the number of people being manipulated, persuaded or deceived but at the prevention of significant harm tied to those effects–in particular, externalities related to political decision-making in all its forms.[216]

Annex B.1:
[Existing] Verification mechanisms

1. To verify States Parties' obligations under Articles 3 and 4, the Agency may take any of the following measures, independently or in combination:
   a. Interview personnel of advanced AI providers or AI hardware manufacturers, with interviews having a narrow scope defined by the Conference;
   b. Intelligence gathering activities, including data and code validation, financial audits, sanity checks, workload classification, analytic output verification, satellite or aerial imagery, measurements of energy consumption, thermal and networking characteristics, open-source information, public and privately disclosed statements from data centers, recordings from cameras at AI data centers, and other forms of human intelligence, signals intelligence or electronic surveillance. The Agency shall disclose these activities to the States Parties in which the activities are being carried out, and at all times ensure that any activity is necessary and proportionate to its objectives, as well as compliant with international human rights law.
   c. Maintaining a chip registry that includes a unique ID and the declared location of chips that are regularly used for advanced AI development or deployment.
   d. Location verification through a network delay-based system in which trusted servers ping the secure ID of chips to measure round-trip latency.
   e. On-site inspections of advanced AI providers' facilities, data centers or hardware manufacturing facilities. These inspections will be carried out in strict compliance with protocols developed by the Agency and approved by the Conference.
2. States Parties shall implement whistleblower programs within their jurisdictions to enable and encourage personnel of advanced AI providers to unveil potential violations of the Articles, to both domestic entities and the Agency.

(1) Arms control and disarmament treaties have historically featured the most elaborate verification regimes, reflecting the high stakes of security agreements. By contrast, agreements in areas like the environment, trade, or human rights often rely more on transparency and reporting, with fewer on-site inspections or enforcement tools. Examples of elaborate verification regimes for high-stakes security agreements include those found in the U.S.–Soviet Strategic Arms Reduction Treaty (START), the Nuclear Non-Proliferation Treaty and the Chemical Weapons Convention.[217] The transboundary challenges presented by AI verification require an arms-like verification mechanism.

(2) As discussed in the commentary to article 5, verification mechanisms for AI need to draw a careful balance between the preservation of privacy and confidentiality on one hand, and effectiveness on the other. Most verification proposals that are both complete and privacy-preserving, like the one included under Annex B.2, still require significant research and development efforts to be at a stage in which they can be tested and adopted. Therefore, while striving to accelerate those efforts as much as possible, the Arti-

cles are realistic and acknowledge that other verification mechanisms may be required in the meantime.

(3) Annex B.1 offers several options for near-term verification, gathered from existing proposals in the AI governance space, especially Baker et al.[218] The first paragraph lists mechanisms that would be under the control of the Agency and divides them into four groups: personnel interviews, intelligence gathering, a chip registry, location verification, and on-site inspections. With the exception of location verification, which has been designed to fit the specific demands of AI governance,[219] these mechanisms are borrowed from analogous international regimes like those concerning nuclear weapons and chemical weapons and are often implemented by the International Atomic Energy Agency and the Organisation for the Prohibition of Chemical Weapons.[220] As such, they have been tested and States would be familiar with them. However, this also means States are aware of their shortcomings.[221] These limitations exhibit the need to increase the amount of resources devoted to verification mechanisms that can be trusted by all States Parties and are designed to fill the specific governance gaps created by advanced AI (see Appendix B.2).

(4) The second paragraph of Annex B.1 establishes whistleblower programs as an additional verification layer. This type of mechanism is crucial to protect employees of AI providers who are closest to the risks and are therefore best placed to warn their governments, the media, the general public, or international organizations that risk thresholds have been crossed.[222] Whistleblowing programs are starting to be adopted in some national laws, like SB53 in California,[223] and companies' transparency frameworks.[224] Contrary to the mechanisms under paragraph 1, these programs would need to be implemented domestically, without direct oversight by the Agency. Therefore, the responsibility to implement them lies on States Parties, which need to create safe channels of communication for whistleblowers to safely share information within their own jurisdiction and with the Agency.

Annex B.2:
[Hardware-enabled] Verification mechanism

To verify States Parties' obligations under Articles 3 and 4, the Agency will take the measures described below.

The Agency shall install, under close supervision of States Parties, one or more neutral data centers (also referred to in this annex as 'verifier's data centers'). It shall also coordinate with States Parties to repurpose or build one or more provers' data centers. All data centers will contain verified hardware, meaning hardware that has been approved by the Agency as complying with the specifications required to carry out the verification mechanism described in this Annex. Hardware specifications must include the ability to demonstrate that tampering has taken place. Agency officials shall routinely inspect both the verifier and the provers' data centers to ensure they comply with the predetermined specifications. States Parties may also continuously monitor both the verifier and the provers' data centers, following protocols developed by the Agency.

The neutral data center shall be under the Agency's ownership and control, but with access to States Parties under the aforementioned Agency protocols. The Agency will be directly accountable to the Conference of States Parties for the administration of the neutral data center. The provers' data centers shall be located in the jurisdiction of States Parties which host hardware used to develop advanced AI.

All verified data centers shall have a digital perimeter, meaning a physical, digital, or hybrid boundary system that prevents data from entering or exiting the enclosed area without being noticed by both the verifier and the prover. The specifications of this digital perimeter shall be defined by the Agency in consultation with States Parties.

Within the digital perimeter of their prover's data center, AI providers shall create cryptographic commitments, meaning the short strings that commit to a piece of data without revealing the actual data, thus allowing the prover to later reveal that data and simultaneously prove that it was the same data that was committed to earlier. The Agency shall also create cryptographic commitments on the computations it has used to verify that an AI system complies with the Articles. The Agency shall publish the results of verification in a way that does not compromise the provers' data.

(1) The mechanism described in Annex B.2 is based on a proposal by Harack and others (2025), which this commentary summarizes and adapts to the Articles' context.[225] The mechanism approximates verifiable confidential computing—computing which perfectly protects an agent's privacy while also providing them with the ability to credibly demonstrate, exactly and completely, that all of their computations adhered to a set of rules under the Articles. This mechanism also incorporates cutting-edge security so that no actor, other than the AI developer or prover, can tell how an AI system was developed or steal their data during any of the planned computational phases. These protections also apply to the verifier's computations.[226] Thus, the mechanism is both secure and verifiable. As discussed above under article 5, this approach would address the tradeoff between transparency and security present in most verification mechanisms, making it more likely that States would be willing to participate in an international governance regime such as the one presented in the Articles.[227]

(2) An additional layer of credibility and reliability is the mechanism's mutual verification system, according to which both the Agency and States Parties have access to each others' data centers to ensure that all the tools and data used within them are true to purpose and not secret attempts to circumvent or undermine the governance system.[228] For this system to be reliable, the neutrality of the Agency's data center is paramount. This neutrality should be safeguarded by both the States Parties and the Agency's host State.[229] In addition, the data center's digital perimeter, key to its neutrality and effectiveness, could be enhanced by a combination of measures such as: (i) largely (or fully) air-gapping the center along with all of the tools and personnel employed, removing the ability to send or receive any kind of remote signals; (ii) verification of all network traffic with specialized hardware, for both the digital perimeter and subcomponents like racks or pods; (iii) a cryptographic boundary composed of data center gateways to enforce encryption on all inputs and outputs, with the use of cryptographic keys for all exchanges, to impede other actors from copying data.[230] States Parties could also consider having multiple neutral facilities in different jurisdictions to provide several independent checks of compliance for some activities.[231]

(3) The mechanism in Annex B.2 could rely on cryptographic commitments about sets of confidential computation or on networking hardware and hardware enclosures to make cryptographic commitments about all traffic. The latter approach could be similarly credible to commitments made via confidential computing, since the Agency and States Parties could inspect data centers to ensure that the verified networking hardware is the only hardware available. There are many unanswered questions about the difficulty, cost, and time involved in deploying a hardware enclosure scheme at scale. However, if these questions were answered, it might be possible to implement both approaches in parallel, thus providing two independent mechanisms—backed by different roots of trust—for making verifiable claims.[232]

(4) The mechanism in Annex B.2 is designed to stand by itself if needed. However, as discussed above under the commentary to article 5, it is important to take into account that it is only one of many proposals for international verification of obligations on AI development[233] and that these proposals (including those under Annex B.1) are not necessarily mutually exclusive. To the contrary, they can act as additional layers of verification or security. Therefore, the Agency should continuously consider using its prerogative under article 5.c to complement the verification mechanism in Annex B.2 with other measures, especially as the technology of verification advances alongside AI development.

Annex C.1:
[Domestic] Enforcement mechanism

When the Agency identifies an AI system that pursues or is capable of any of the prohibited uses defined in article 3, or presents any of the risks or capabilities listed in Annex A, it shall notify any States Parties with jurisdiction over the violation. These States Parties shall execute the enforcement mechanism described below.

All chips used for AI development or deployment are subject to an authorization to operate, which is given by the State Party in which an AI provider is located. This authorization will be automatically renewed periodically (potentially every few minutes) unless a violation of article 3 is verified by the Agency. The authorization will consist of a certificate that is cryptographically signed by the State Party's private key. Firmware running on the chip's security processor will verify that the signature is from a private key that coincides with the State Party's public key, also stored in the firmware binary. The identification number of each cryptographic signature will be securely stored in the chip once it has received a new authorization. If a new authorization does not meet these conditions, the chip will refuse to complete any new computations until the State Party provides a new valid authorization.

If a violation of article 3 is verified by the Agency, the relevant State Party shall not extend a new authorization for the chips being used to run the AI system found to be in violation. At the start of the next clock cycle, those chips will detect a lack of authorization and refuse to complete any computations until the State Party provides a new authorization.

All chips used for AI development must be placed in a tamper-proof secure enclosure along with an auxiliary guarantee processor. All of the chip's memory and other components will also be contained within the tamper-proof secure enclosure. If an attempt is made to tamper with the enclosure, the chip will automatically wipe the guarantee processor's firmware and microscopic fuses on the AI chip will be activated to render the chip permanently inoperable.

States Parties shall install a firmware update in guarantee processors as regularly as needed to improve the processor's security or implement changes in the conditions for authorizations.

> *Commentaries for annex C.2 are the same as for annex C.1.*

Annex C.2:
[International] Enforcement mechanism

When the Agency identifies an AI system that pursues or is capable of any of the prohibited uses defined in article 3, or presents any of the risks or capabilities listed in Annex A, it shall execute the enforcement mechanism described below.

All chips used for AI development or deployment are subject to an authorization to operate, which is given by the Agency. This authorization will be automatically renewed periodically (potentially every few minutes) unless a violation of article 3 is verified by the Agency. The authorization will consist of a certificate that is cryptographically signed by the Agency's private key. Firmware running on the chip's security processor will verify that the signature is from a private key that coincides with the Agency's public key, also stored in the firmware binary. The identification number of each cryptographic signature will be securely stored in the chip once it has received a new authorization. If a new authorization does not meet these conditions, the chip will refuse to complete any new computations until the Agency provides a new valid authorization.

If a violation of article 3 is verified by the Agency, it will not extend a new authorization for the chips being used to run the AI system found to be in violation. At the start of the next clock cycle, those chips will detect a lack of authorization and refuse to complete any computations until the Agency provides a new authorization.

All chips used for AI development must be placed in a tamper-proof secure enclosure along with an auxiliary guarantee processor. All of the chip's memory and other components will also be contained within the tamper-proof secure enclosure. If an attempt is made to tamper with the enclosure, the chip will automatically wipe the guarantee processor's firmware and microscopic fuses on the AI chip will be activated to render the chip permanently inoperable.

The Agency may require States Parties to install a firmware update in guarantee processors as regularly as needed to improve the processor's security or implement changes in the conditions for authorizations.

(1) The enforcement mechanism in Annex C consists of an offline cryptographic authorization measure, based primarily on the proposal made in Petrie (2024) and complemented by the work of Aarne and others (2024), Kulp and others (2024) and Petrie and others (2025).[234] The mechanism consists of an offline licensing framework executed through flexible hardware-enabled guarantees. The main conditions for this mechanism to function are: (i) the installation of a guarantee processor in chips used for AI development; (ii) the embedding of a public key, belonging to the enforcer–whether a State Party or the Agency, depending on whether the mechanism is implemented domestically or internationally– in the guarantee processor's firmware; (iii) tamper-secure enclosures for each chip, which include microscopic fuses to disable chips; and (iv) an enforcer-managed system which automatically produces cryptographic signatures acting as authorization for chips to complete computations. Under article 4, States Parties must guarantee that AI chips produced within their jurisdiction meet these conditions..

(2) The mechanism is built around authorizations or licenses. Each authorization is a small file that contains: the number of allowed clock cycles per authorization, which chip the authorization is for, and an authorization identification number so that the chip can check that the authorization has not been used before. Authorizations are cryptographically signed using the private key(s) of the enforcer–whether a State Party or the Agency– so that the AI chip can verify that the authorization is authentic.[235] As discussed under the commentary for Article 6, Annex C provides two options for enforcement–one domestic and one international. Depending on the mechanism that is ultimately selected, the actor in charge of providing authorizations and implementing the enforcement mechanism would be the State Party in which a violation has been detected or the Agency.

(3) The cryptographic authorization of computations by each chip revolves around clock cycles–the length of time in which an authorization is valid before a chip requires a new authorization to continue executing computations. As explained by Petrie, clock cycles offer several advantages over other metrics because they are the closest proxy to chip usage time, they are able to enable or disable chips without fine-grained usage controls (which is less intrusive for AI providers), chips are easily able to keep track of clock cycles, and clock cycles can have an important secondary function in detecting tampering.[236] The clock cycle in Annex C is meant to be defined by the enforcer–whether a State Party or the Agency–but it should be short enough (perhaps every few minutes) to limit the chances that an AI system could be misused or cause harm autonomously if a violation of article 3 is detected by the Agency.

(4) The offline cryptographic authorization mechanism in Annex C has several additional advantages. First, the cryptographic signatures and keys that are only available to the enforcer–whether a State Party or the Agency–reduce the chances that AI providers, States Parties or third party actors might be able to misuse chips in a way that breaches obligations under the Articles. Second, this enforcement mechanism has space for corrective actions–while other proposals for enforcement rely on a permanent disabilitation of chips or other forms of hardware, Annex C leaves open the window for chips to be re-authorized to execute computations, if AI providers correct their actions or are able to demonstrate that the Agency's alleged verification of a violation was inaccurate. Third, the Annex counts with an additional layer of security through tamper-proof enclosures, preventing the continued operation of chips that have been modified to bypass the Agency's verification or the enforcement mechanism. Lastly, to add yet another layer of security, Annex C gives the enforcer–whether a State Party or the Agency– the mandate to update the firmware of guarantee processors in chips to improve their security or change the conditions (the "ruleset") for new authorizations to be valid.[237]

Bibliography

1 Epoch AI, *AI Benchmarking Hub*, epoch.ai (last updated Oct. 1, 2025), https://epoch.ai/benchmarks; Thomas Kwa et al., *Measuring AI Ability to Complete Long Tasks*, arXiv:2503.14499 (Mar. 18, 2025), https://arxiv.org/abs/2503.14499.

2 See The Elders, *The Elders urge global co-operation to manage risks and share benefits of AI*, The Elders (May 31, 2023), https://theelders.org/news/elders-urge-global-co-operation-manage-risks-and-share-benefits-ai; Stanford Inst. for Human-Centered Artificial Intelligence, *Policy and Governance*, in Artificial Intelligence Index Report 2025 (2025), https://hai.stanford.edu/ai-index/2025-ai-index-report/policy-and-governance.

3 For a clear example of advanced AI capabilities being used by malicious actors, see *Disrupting the first reported AI-orchestrated cyber espionage campaign*, Anthropic (Nov. 13, 2025), https://www.anthropic.com/news/disrupting-AI-espionage. For more examples of AI-related incidents, see MIT, *MIT AI Incident Tracker* (2025), https://airisk.mit.edu/ai-incident-tracker; OECD, *OECD AIM: AI Incidents and Hazards Monitor*, OECD.AI (2025), https://oecd.ai/en/incidents; Ctr. for Sec. & Emerging Tech., *AI Incident Database*, https://incidentdatabase.ai/; *AI Risk Explorer*, https://www.airiskexplorer.com/; Caroline Jeanmaire & Sam Boger, *AI Incidents Are Rising. It's Time for the United States to Build Playbooks for When AI Fails.*, The Future Soc'y (Nov. 12, 2025), https://thefuturesociety.org/us-ai-incident-response/.

4 See, for example, Anthropic, *System Card: Claude Opus 4 & Claude Sonnet 4* (2025), parts 3 and 4; Antonio Regalado, *Microsoft Says AI Can Create Zero-Day Threats in Biology*, MIT Tech. Rev. (Oct. 2, 2025), https://www.technologyreview.com/2025/10/02/1124767/microsoft-says-ai-can-create-zero-day-threats-in-biology/.

5 Oliver Guest, *Prospects for AI Safety Agreements Between Countries*, Rethink Priorities, https://rethinkpriorities.org/research-area/prospects-for-ai-safety-agreements-between-countries/.

6 U.N. High-level Advisory Body on Artificial Intelligence, *Governing AI for Humanity: Final Report*, ¶ 206 (2024); Thomas A. Birkland & Kathryn L. Schwaeble, *Agenda Setting and the Policy Process: Focusing Events*, Oxford Rsch. Encyclopedia of Pol. (June 25, 2019), https://oxfordre.com/politics/display/10.1093/acrefore/9780190228637.001.0001/acrefore-9780190228637-e-165.

7 Tolga Bilge and others, *A Treaty to Manage AI Risk*, AITreaty.org, https://aitreaty.org/; A Narrow Path, *Phase 0*, in A Narrow Path (Oct. 2, 2024), https://www.narrowpath.co/phase-0; https://arxiv.org/pdf/2503.18956; https://www.cigionline.org/static/documents/Harris-Deb-Paper302.pdf; If Anyone Builds It, Everyone Dies, *A Tentative Draft of a Treaty, with Annotations*, https://ifanyonebuildsit.com/treaty.

8 The Am. L. Inst., *Statement on Essential Human Rights*, The American Law Institute (June 2, 2020), https://www.ali.org/news/articles/statement-essential-human-rights.

9 Emanuel Adler, *The Emergence of Cooperation: National Epistemic Communities and the International Evolution of the Idea of Nuclear Arms Control*, 46 INT'L ORG. 101 (1992). Also see Peter M. Haas, *Introduction: Epistemic Communities and International Policy Coordination*, 46 Int'l Org. 1 (1992).

10 José Jaime Villalobos, Matthijs Maas and Christoph WInter, *International Catastrophe Law: from Nuclear Weapons to Advanced AI* (Cambridge University Press, forthcoming 2026).

11 See list of international initiatives in *Global Governance of AI*, global-governance.ai.

12 Danae Azaria, *The Indirect Legal Effects of Non-Binding Agreements* (Nov. 7, 2023), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4604896.

13 Org. for Econ. Co-operation & Dev., *OECD AI Principles*, OECD.AI (2025), https://oecd.ai/en/ai-principles.

14 *The Bletchley Declaration*, Gov.UK (Nov. 1, 2023), https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration.

15 *The Seoul Declaration*, Gov.UK (May 21, 2024), https://www.gov.uk/government/publications/seoul-declaration-for-safe-innovative-and-inclusive-ai-ai-seoul-summit-2024.

16 Recommendation on the Ethics of Artificial Intelligence, UNESCO (May 16, 2023), https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence.

17 Ministry of Internal Affs. & Commc'ns of Japan, *Hiroshima AI Process*, https://www.soumu.go.jp/hiroshimaaiprocess/en/index.html.

18 Int'l Chamber of Com., *Harmonised AI standards to reduce fragmented global rules*, (July 11, 2025), https://iccwbo.org/news-publications/policies-reports/harmonised-ai-standards-to-reduce-fragmented-global-rules/.

19 Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (opened for signature 5 September 2024, not yet in force) CETS No 225.

20 José Jaime Villalobos, Matthijs Maas and Christoph WInter, *International Catastrophe Law: from Nuclear Weapons to Advanced AI* (Cambridge University Press, forthcoming 2026).

21 Dep't for Sci., Innovation & Tech. & AI Safety Inst., *International AI Safety Report 2026* (2026).

22 See, for example, MIT, *MIT AI Incident Tracker* (2025), https://airisk.mit.edu/ai-incident-tracker; OECD, *OECD AIM: AI Incidents and Hazards Monitor*, OECD.AI (2025), https://oecd.ai/en/incidents; Ctr. for Sec. & Emerging Tech., *AI Incident Database*, https://incidentdatabase.ai/; *AI Risk Explorer*, https://www.airiskexplorer.com/.

23 Steven J. Hoffman et al., *International treaties have mostly failed to produce their intended effects*, 119 PNAS (2022).

24 Matthijs Maas and José Jaime Villalobos, *International AI institutions: A literature review of models, examples, and proposals*, Inst. for L. & AI (Sept. 2023), https://law-ai.org/international-ai-institutions/; U.N. High-level Advisory Body on Artificial Intelligence, *Governing AI for Humanity: Final Report* (2024).

25 U.N. Glob. Digit. Compact, *AI Panel and Dialogue*, https://www.un.org/global-digital-compact/en/ai.

26 Stanford Inst. for Human-Centered Artificial Intelligence, *Artificial Intelligence Index Report 2025* (2025), chs 5, 7; David Rolnick et al., *Tackling Climate Change with Machine Learning*, 55 ACM Computing-*Surveys* 1 (2022).

27 While the United States, and to a certain extent China, concentrate most of the compute necessary to develop advanced AI systems, several other States that already have a significant number of data centers are only likely to continue accumulating compute. See Paul Mozur & Adam Satariano, *The Global A.I. Divide*, N.Y. Times (June 23, 2025), https://www.nytimes.com/interactive/2025/06/23/technology/ai-computing-global-divide.html

28 José Jaime Villalobos, Matthijs Maas and Christoph WInter, *International Catastrophe Law: from Nuclear Weapons to Advanced AI* (Cambridge University Press, forthcoming 2026); Matthijs Maas and José Jaime Villalobos, *International AI institutions: A literature review of models, examples, and proposals*, Inst. for L. & AI (Sept. 2023), https://law-ai.org/international-ai-institutions/; Duncan Cass-Beggs et al., *Framework Convention on Global AI Challenges*, Ctr. for Int'l Governance Innovation (June 25, 2024), https://www.cigionline.org/publications/framework-convention-on-global-ai-challenges/.

29 See *Global Digital Compact*, G.A. Res. 79/1, Annex, U.N. Doc. A/RES/79/1, at 3 (Sept. 22, 2024), Objective 5.

30 Anthony Aguirre, *Keep the Future Human* https://keepthefuturehuman.ai/wp-content/uploads/2025/03/Keep_the_Future_Human__AnthonyAguirre__5March2025.pdf

31 José Jaime Villalobos, Matthijs Maas and Christoph WInter, *International Catastrophe Law: from Nuclear Weapons to Advanced AI* (Cambridge University Press, forthcoming 2026).

32 Matthijs Maas and José Jaime Villalobos, *International AI institutions: A literature review of models, examples, and proposals*, Inst. for L. & AI (Sept. 2023), https://law-ai.org/international-ai-institutions/.

33 Duncan Cass-Beggs et al., *Framework Convention on Global AI Challenges*, Ctr. for Int'l Governance Innovation (June 25, 2024), https://www.cigionline.org/publications/framework-convention-on-global-ai-challenges/.

34 For an explanation of each of these building blocks and an interactive tool showing how current international efforts relate to each one, see *Global Governance of AI*, global-governance.ai.

35 Regulation (EU) 2024/1689, 2024 O.J. (L 2024/1689) 1, art 3. For commentary on the definitions of "AI system" and "systemic risk" in the EU AI Act, see *Recitals - AI Act*, AI-ACT-LAW.EU, https://ai-act-law.eu/recital/, recitals 12, 110 (with recital 110 indicating: "General-purpose AI models could pose systemic risks which include, but are not limited to, any actual or reasonably foreseeable negative effects in relation to major accidents, disruptions of critical sectors and serious consequences to public health and safety; any actual or reasonably foreseeable negative effects on democratic processes, public and economic security; the dissemination of illegal, false, or discriminatory content. Systemic risks should be understood to increase with model capabilities and model reach, can arise along the entire lifecycle of the model, and are influenced by conditions of misuse, model reliability, model fairness and model security, the level of autonomy of the model, its access to tools, novel or combined modalities, release and distribution strategies, the potential to remove guardrails and other factors.")

36 S.B. 53, 2025–2026 Leg., Reg. Sess. (Cal. 2025), sec 22757.11(c)(1).

37 Samantha Beeson, *Sovereignty* (April 2011), Max Planck Encyclopedia of Public International Law, https://opil.ouplaw.com/display/10.1093/law:epil/9780199231690/law-9780199231690-e1472.

38 International Covenant on Economic, Social and Cultural Rights, Dec. 16, 1966, 993 U.N.T.S. 3, art 15.

39 Michael C. Horowitz and Lauren A. Khan, *Nuclear Non-Proliferation is the Wrong Framework for AI Governance*, AI Frontiers (2025), https://aifrontiersmedia.substack.com/p/nuclear-non-proliferation-is-the

40 Yuan Yi Zhu, *Artificial intelligence, export controls, and great power competition*, Expert Essentials (2025), https://doi.org/10.1093/9780198972877.003.0038.

41 Dan Hendrycks, Eric Schmidt and Alexandr Wang, *Superintelligence Strategy: Expert Version*, arXiv:2503.05628 (Apr. 14, 2025), https://arxiv.org/abs/2503.05628; Sam Winter-Levy, *The AI Export Dilemma: Three Competing Visions for U.S. Strategy*, Carnegie Endowment for International Peace (Dec. 13, 2024), https://carnegieendowment.org/research/2024/12/ai-artificial-intelligence-export-united-states?lang=en

42 Epoch, *Trends in Machine Learning*, epoch.ai (last updated Oct. 1, 2025), https://epoch.ai/trends#hardware.

43 There are some indications that China, for example, might be able to reach this goal. See Craig Smith, *China's AI Self-Sufficiency Push Is Challenging U.S. Dominance*, FORBES (Sept. 2, 2025), https://www.forbes.com/sites/craigsmith/2025/09/02/chinas-ai-self-sufficiency-push-is-challenging-us-dominance/.

44 Helen Toner, *"Nonproliferation" is the wrong frame for the dangers of very powerful future AI*, Substack (Sept. 28, 2025), https://helentoner.substack.com/p/nonproliferation-is-the-wrong-approach.

45 For a discussion on the origins of limitations to national sovereignty and self-determination in the context of human rights, see Daniel Moeckli et al., *International Human Rights Law* (Oxford University Press, 2014), ch 1, 98-101, 345-348.

46 For examples, see International Covenant on Civil and Political Rights (signed 16 December 1966, entered into force 23 March 1976) 999 UNTS 171 (ICCPR), arts 4, 12, 18, 19, 21, 22; International Covenant on Economic, Social and Cultural Rights (adopted 16 December 1966, entered into force 3 January 1976) 993 UNTS 3 (ICESCR), art 4; General Agreement on Tariffs and Trade (signed 30 October 1947, entered into force 1 January 1948) 55 UNTS 187 (GATT), arts XX, XXI; General Agreement on Trade in Services, annexed to the Marrakesh Agreement Establishing the World Trade Organization (signed 15 April 1994, entered into force 1 January 1995) 1869 UNTS 183 (GATS), arts XIV, XIV bis; Agreement on Trade-Related Aspects of Intellectual Property Rights, Annex 1C to the Agreement establishing the World Trade Organization (signed 15 April 1994, entered into force 1 January 1995) 1869 UNTS 299 (TRIPS), arts 13, 27, 30, 73.
An exception to this pattern of limitations or derogations is that of rules that fall under the category of *jus cogens*–many of which may apply to the mitigation of risks from AI development. See José Jaime Villalobos, Matthijs Maas and Christoph WInter, *International Catastrophe Law: from Nuclear Weapons to Advanced AI* (Cambridge University Press, forthcoming 2026), ch 4.

47 U.N. Human Rights Comm., General Comment No. 34, para 32 n.4, U.N. Doc. CCPR/C/GC/34 (2011).

48 Sacha Alanoca et al, *Comparing Apples to Oranges: A Taxonomy for Navigating the Global Landscape of AI Regulation,* arXiv:2505.13673 (May 19, 2025), https://arxiv.org/abs/2505.13673.

49 See, for example, Montreal Protocol, CITES, WHC.

50 Jonas Schuett et al, *From Principles to Rules: A Regulatory Approach for Frontier AI*, arXiv:2407.07300 (Jul. 10, 2024), https://arxiv.org/abs/2407.07300.

51 Leonardo Berti, Flavio Giorgi and Gjergji Kasneci, *Emergent Capabilities in Large Language Models: A Survey*, arXiv:2503.05788 (Feb. 28, 2025), https://arxiv.org/abs/2503.05788.

52 For examples of proposals using FLOP as the anchor for a legal obligation, see A Narrow Path, *Phase 0*, in A Narrow Path (Oct. 2, 2024), https://www.narrowpath.co/phase-0; Rebecca Scholefield, Samuel Martin and Otto Barten, *International Agreements on AI Safety: Reviews and Recommendations for a Conditional AI Safety Treaty* https://arxiv.org/pdf/2503.18956 (Mar. 18, 2025); Anthony Aguirre, *Keep the Future Human* https://keepthefuturehuman.ai/wp-content/uploads/2025/03/Keep_the_Future_Human__AnthonyAguirre__5March2025.pdf, 13.

53 Moshe Hoffman et al., *Why norms are categorical* (Feb. 17, 2018), Harvard University, https://www.tse-fr.eu/sites/default/files/TSE/documents/sem2018/eco_pub/hoffman.pdf.

54 Rebecca Crootoff, "Jurisprudential Space Junk: Treaties and New Technologies" in Chiara Giorgetti and Natalie Klein (eds), *Resolving Conflicts in the Law*, Brill (2019), ch 7.

55 Jonas Schuett et al, *Surveys on thresholds for advanced Ai systems*, Centre for the Governance of AI (2025), https://aigi.ox.ac.uk/wp-content/uploads/2025/08/Survey_on_thresholds_for_advanced_AI_systems_1.pdf; Deepika Raman et al., *Intolerable Risk Threshold Recommendations for Artificial Intelligence*, UC Berkley Center for Long-Term Cybersecurity (2025, https://cltc.berkeley.edu/wp-content/uploads/2025/02/Intolerable-Risk-Threshold-Recommendations-for-Artificial-Intelligence.pdf, 16-20.

56 See, for example, Anthony Aguirre, *Keep the Future Human* https://keepthefuturehuman.ai/wp-content/uploads/2025/03/Keep_the_Future_Human__AnthonyAguirre__5March2025.pdf, 44. Some AI corporations have also developed their own risk tiers as part of their 'responsible scaling policies', under which certain voluntary safety measures are taken according to the level of risk; see, for example, Anthropic, *Anthropic's Responsible Scaling Policy* (Sep. 19, 2023), https://www.anthropic.com/news/anthropics-responsible-scaling-policy; OpenAI, *Preparedness Framework* (Apr. 15, 2025), https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf; Allan Dafoe et al., *Updating the Frontier Safety Framework*, Google Deepmind Blog (Feb. 4, 2025), https://deepmind.google/blog/updating-the-frontier-safety-framework/.

57 See Int'l L. Comm'n, *Draft Articles on the Protection of Persons in the Event of Disasters, with commentaries*, *in* Rep. of the Int'l L. Comm'n on the Work of Its Sixty-Eighth Session, UN GAOR, 71st Sess., Supp. No. 10, at 19, UN Doc. A/71/10 (2016), art 9; Samantha Besson, *Due Diligence in International Law* (Brill 2023); *Request for an Advisory Opinion Submitted by the Comm'n of Small Island States on Climate Change & Int'l Law (COSIS)*, ¶ 441(3)(c) (ITLOS Case No. 31, Advisory Opinion of May 21, 2024); Antonio Coco & Talita Dias, *"Handle with Care": Due Diligence Obligations in the Employment of AI Technologies*, in *Research Handbook on Warfare and Artificial Intelligence* (Robin Geiß & Henning Lahmann eds., Edward Elgar 2024) (discussing due diligence obligations on the design, development, and deployment of AI technologies, as grounded in due diligence standards of conduct found in different positive protective obligations, from *Corfu Channel* and the no-harm principles, to positive duties under international humanitarian and human rights law); *Pulp Mills on the River Uruguay (Arg. v. Uru.)*, [2010] I.C.J. Rep. 14, ¶ 204 ("'the obligation to protect and preserve, under Article 41 (a) of the Statute, has to be interpreted in accordance with a practice, which in recent years has gained so much acceptance among States that it may now be considered a requirement under general international law to undertake an environmental impact assessment where there is a risk that the proposed industrial activity may have a significant adverse impact in a transboundary context, in particular, on a shared resource"); *Certain Activities Carried out by Nicaragua in the Border Area (Costa Rica v. Nicar.) & Construction of a Road in Costa Rica along the San Juan River (Nicar. v. Costa Rica)*, [2015] I.C.J. Rep. 665, ¶ 104; *Responsibilities & Obligations of States Sponsoring Persons & Entities with Respect to Activities in the Area (Seabed Advisory Opinion)*, [2011] ITLOS Rep. 10, ¶ 145 (Feb. 1, 2011), As discussed in: Stephen Townley, *The Rise of Risk in International Law*, 18 CHI. J. INT'L L. 594, 598 (2018).

58 Anthony Aguirre, *Keep the Future Human* https://keepthefuturehuman.ai/wp-content/uploads/2025/03/Keep_the_Future_Human__AnthonyAguirre__5March2025.pdf, 44; Narrow Path, *Phase 0*, in A Narrow Path (Oct. 2, 2024), https://www.narrowpath.co/phase-0; Rebecca Scholefield, Samuel Martin and Otto Barten, *International Agreements on AI Safety: Reviews and Recommendations for a Conditional AI Safety Treaty* https://arxiv.org/pdf/2503.18956 (Mar. 18, 2025); more generally, see Matthijs Maas and José Jaime Villalobos, *International AI institutions: A literature review of models, examples, and proposals*, Inst. for L. & AI (Sept. 2023), https://law-ai.org/international-ai-institutions/, 20-25.

59 Lennart Heim et al., *Governing Through the Cloud: The Intermediary Role of Compute Providers in AI Regulation* (Mar. 13, 2024), https://www.governance.ai/research-paper/governing-through-the-cloud; Girish Sastry et al., *Computing Power and the Governance of Artificial Intelligence* (Feb. 14, 2024), https://arxiv.org/abs/2402.08797.

60 Obligations to "take measures" or "take all necessary measures" are not uncommon in various fields of international law. The Treaty on the Prohibition of Nuclear Weapons, for instance, indicates that "Each State Party shall take all necessary legal, administrative and other measures, including the imposition of penal sanctions, to prevent and suppress any activity prohibited to a State Party under this Treaty undertaken by persons or on territory under its jurisdiction or control." Treaty on the Prohibition of Nuclear Weapons, July 7, 2017, 3370 U.N.T.S. (registration No. 56487), art 5. For more examples, see Convention on the Rights of the Child, Nov. 20, 1989, 1577 U.N.T.S. 3, arts 11, 24, 34; Convention on the Elimination of All Forms of Discrimination against Women, Dec. 18, 1979, 1249 U.N.T.S. 13, art 3; Stockholm Convention on Persistent Organic Pollutants, May 22, 2001, 2256 U.N.T.S. 119, art 5; Convention on Cluster Munitions, Dec. 3, 2008, 2688 U.N.T.S. 39, art 9.

61 For examples of the types of measures one could take to fulfill this obligation, see Onni Aarne, Tim Fist & Caleb Withers, *Secure, Governable Chips: Using On-Chip Mechanisms to Manage National Security Risks from AI & Advanced Computing* (Ctr. for a New Am. Sec. 2024), https://s3.us-east-1.amazonaws.com/files.cnas.org/documents/CNAS-Report-Tech-Secure-Chips-Jan-24-finalb.pdf; James Petrie and Ooni Aarne, *Technical Options for Flexible Hardware-Enabled Guarantees* (June 2025), https://arxiv.org/abs/2506.03409; Miles Brundage et al., *Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims* (Apr. 21, 2020), https://arxiv.org/abs/2004.07213.

62 Dep't for Sci., Innovation & Tech. & AI Safety Inst., *International AI Safety Report 2026* (2026), 226.

63 Mauricio Baker et al., *Verifying International Agreements on AI: Six Layers of Verification for Rules on Large-Scale AI Development and Deployment* (RAND Corp. Working Paper No. WR-A4077-1, 2025), https://www.rand.org/pubs/working_papers/WRA4077-1.html.

64 Kenneth W. Abbott, *Trust But Verify: The Production of Information in Arms Control Treaties and Other International Agreements*, 26 CORNELL INT'L L.J. 1 (1993).

65 Akash R. Wasil et al., *Verification methods for international AI agreements* (Aug. 28, 2024), https://arxiv.org/abs/2408.16074.

66 Andrew J. Coe & Jane Vaynman, *Why Arms Control Is So Rare*, 114 AM. POL. SCI. REV. 342 (2020).

67 Ben Harack et al., *Verification for International AI Governance* (AI Governance Initiative July 2025), https://aigi.ox.ac.uk/wp-content/uploads/2025/07/Verification_for_International_AI_Governance.pdf, 30-31, 64; Aaron Scher, David Abecassis, Peter Barnett, & Brian Abeyta, *An International Agreement to Prevent the Premature Creation of Artificial Superintelligence* (Nov. 13, 2025), https://arxiv.org/pdf/2511.10783, 50.

68 Mauricio Baker et al., *Verifying International Agreements on AI: Six Layers of Verification for Rules on Large-Scale AI Development and Deployment* (RAND Corp. Working Paper No. WR-A4077-1, 2025), https://www.rand.org/pubs/working_papers/WRA4077-1.html; Aaron Scher & Lisa Thiergart, *Mechanisms to Verify International Agreements About AI Development* (Mach. Intel. Rsch. Inst., Nov. 27, 2024), https://techgov.intelligence.org/research/mechanisms-to-verify-international-agreements-about-ai-development; Akash R. Wasil et al., *Verification methods for international AI agreements* (Aug. 28, 2024), https://arxiv.org/abs/2408.16074.

69 Steven J. Hoffman et al., *International treaties have mostly failed to produce their intended effects*, 119 PNAS (2022).

70 See Onni Aarne, Tim Fist & Caleb Withers, *Secure, Governable Chips: Using On-Chip Mechanisms to Manage National Security Risks from AI & Advanced Computing* (Ctr. for a New Am. Sec. 2024), https://s3.us-east-1.amazonaws.com/files.cnas.org/documents/CNAS-Report-Tech-Secure-Chips-Jan-24-finalb.pdf; Lennart Heim et al., *Governing Through the Cloud: The Intermediary Role of Compute Providers in AI Regulation* (Mar. 13, 2024), https://www.governance.ai/research-paper/governing-through-the-cloud; Hussain Al-Aqrabi & Richard Hill, *A Scalable Model for Secure Multiparty Authentication* (Jan. 11, 2019), https://arxiv.org/abs/1901.03056.

71 See, for example, James Petrie et al., *Flexible Hardware-Enabled Guarantees for AI Compute* (June 18, 2025), https://arxiv.org/abs/2506.15093, 7 (recommending permanent disablement but only in response to physical tampering that could circumvent the governance mechanism).

72 INT'L ORG. FOR STANDARDIZATION & INT'L ELECTROTECHNICAL COMM'N, *Safety aspects — Guidelines for their inclusion in standards* (ISO/IEC Guide 51:2014) (2014); INT'L ORG. FOR STANDARDIZATION & INT'L ELECTROTECHNICAL COMM'N, *Guide to the development and inclusion of aspects of safety in International Standards for medical devices* (ISO/IEC Guide 63:2019) (2019); INT'L ORG. FOR STANDARDIZATION, *Medical devices - Application of risk management to medical devices* (ISO 14971:2019) (2019); INT'L ORG. FOR STANDARDIZATION, *Risk management — Guidelines* (ISO 31000:2018) (2018); INT'L ORG. FOR STANDARDIZATION, *Risk management — Vocabulary* (ISO 31073:2022) (2022) [formerly known as Guide 73]; INT'L ORG. FOR STANDARDIZATION & INT'L ELECTROTECHNICAL COMM'N, *Information technology — Artificial intelligence — Risk management* (ISO/IEC 23894:2023) (2023); INT'L ORG. FOR STANDARDIZATION & INT'L ELECTROTECHNICAL COMM'N, *Information technology — Artificial intelligence — Management system* (ISO/IEC 42001:2023) (2023). This commentary purposely avoids referencing ISO 31000/23894, which mostly consists of recommendations, not requirements, for organizational risk management rather than safety-focused risk management.
Note that domestic efforts to provide input to ISO/IEC JTC 1/SC 42 are currently under way. For example, the United States' National Institute of Standards and Technology is developing a zero draft on AI testing, evaluation, verification, and validation as input to that committee. NAT'L INST. OF STANDARDS & TECH., *Outline: Proposed Zero Draft for a Standard on AI Testing, Evaluation, Verification, and Validation (TEVV)* (July 15, 2025), https://www.nist.gov/system/files/documents/2025/07/15/Outline_%20Proposed%20Zero%20Draft%20for%20a%20Standard%20on%20AI%20TEVV-for-web.pdf. Also see the United Kingdom's efforts to develop quantitative safety guarantees for AI through the Advanced Research + Innovation Agency's 'Safeguarded AI' program. ADVANCED RSCH. + INNOVATION AGENCY, *Safeguarded AI Funding*, https://www.aria.org.uk/opportunity-spaces/mathematics-for-safe-ai/safeguarded-ai/funding.

73 INT'L ORG. FOR STANDARDIZATION, *Risk management — Guidelines* (ISO 31000:2018) (2018), 6.5.2, 6.3.4.

74 Although note that ISO 14971 considers benefits because of the nature of medical care, which should be avoided in the case of AI.

75 ORG. FOR ECON. CO-OP. & DEV., *Launch of the Hiroshima AI Process Reporting Framework*, https://www.oecd.org/en/events/2025/02/launch-of-the-hiroshima-ai-process-reporting-framework.html; EUROPEAN COMM'N, *The General-Purpose AI Code of Practice* (July 10, 2025), https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai.

76 EUROPEAN COMM'N, *The General-Purpose AI Code of Practice* (July 10, 2025), https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai, 8.1.

77 See, for example, Anthropic, *The Long-Term Benefit Trust*, ANTHROPIC (Oct. 12, 2023), https://www.anthropic.com/news/the-long-term-benefit-trust.

78 Jonas Schuett, Ann-Katrin Reuel & Alexis Carlier, *How to Design an AI Ethics Board*, 5 AI & ETHICS 863 (2025).

79 See, for example, OpenAI, *OpenAI Charter*, https://openai.com/charter/ ("OpenAI's mission is to ensure that artificial general intelligence (AGI)—by which we mean highly autonomous systems that outperform humans at most economically valuable work—benefits all of humanity.")

80 See, for example, Anthropic, *The Long-Term Benefit Trust*, Anthropic (Oct. 12, 2023), https://www.anthropic.com/news/the-long-term-benefit-trust.

81 Siméon Campos et al., *A Frontier AI Risk Management Framework: Bridging the Gap Between Current AI Practices and Established Risk Management*, arXiv:2502.06656 (Feb. 10, 2025), https://arxiv.org/abs/2502.06656; Ben Robinson et al., *Why Frontier AI Safety Frameworks Need to Include Risk Governance*, The Centre for Long-Term Resilience (Feb. 5, 2025), https://www.longtermresilience.org/reports/frontier-ai-safety-frameworks-need-to-include-risk-governance; Jonas Schuett, *Three lines of defense against risks from AI*, 40 AI & Soc. 493 (2025).

82 See *European Cent. Bank, Strong risk culture – sounds bank (Feb. 15, 2023),* https://www.bankingsupervision.europa.eu/press/supervisory-newsletters/newsletter/2023/html/ssm.nl230215_3.en.html.

83 Hannah West, *Speak Up Culture — What Is It, and Why Is It Important? GO*ODCOURSE (Feb. 10, 2023), https://www.goodcourse.co/post/what-is-speak-up-culture.

84 See Corinna Ewelt-Knauer, Anja Schwering & Sandra Winkelmann, *Doing Good by Doing Bad: How Tone at the Top and Tone at the Bottom Impact Performance-Improving Noncompliant Behavior*, 175 J. Bus. Ethics 609 (2022), https://link.springer.com/article/10.1007/s10551-020-04647-6.

85 David Manheim, *Building a Culture of Safety for AI: Perspectives and Challenges* (June 26, 2023) (unpublished manuscript), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4491421.

86 EUROPEAN COMM'N, *The General-Purpose AI Code of Practice* (July 10, 2025), https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai, 8.3.

87 The beta version of OpenAI's Preparedness Framework, for example, included references to practices like "safety drills" (these have now been "deprioritized"). OpenAI, *Preparedness Framework (Beta)* (Dec. 18, 2023), https://cdn.openai.com/openai-preparedness-framework-beta.pdf.

88 Some industries operate under rolling mandatory reporting of characteristics that could affect their ability to behave safely. In the most advanced biosafety for instance, personnel needs to self-report "medical matters, prescription and over-the-counter medications used, alcohol abuse, legal actions, public record court actions (eg, separation, divorce, lawsuit), financial problems or concerns, or any other activity that may influence the staff member's day-to-day reliability". Peers and supervisors also need to report any elements that might indicate that an individual's day-to-day reliability has been affected. For an example of this type of standard in biosecurity, see Jacki J. Higgins et al., *Implementation of a Personnel Reliability Program as a Facilitator of Biosafety and Biosecurity Culture in BSL-3 and BSL-4 Laboratories*, 11 Biosecur Bioterror 130 (2013), https://pmc.ncbi.nlm.nih.gov/articles/PMC3689182/.

89 See Apollo Rsch., *AI Behind Closed Doors: a Primer on The Governance of Internal Deployment* (Apr. 17, 2025), https://www.apolloresearch.ai/research/ai-behind-closed-doors-a-primer-on-the-governance-of-internal-deployment-apollo-research/.

90 See Jimmy Farrell, *Learning from History: GPAI serious incident reporting*, Pour Demain (Oct. 18, 2024), https://www.pourdemain.ngo/en/post/learning-from-history-gpai-serious-incident-reporting.

91 See Kyra Dempsey, *Why You've Never Been In A Plane Crash*, Asterisk, Feb. 2024, at 5, https://asteriskmag.com/issues/05/why-you-ve-never-been-in-a-plane-crash.

92 SB53, California's AI transparency bill, already establishes protections for whistleblowers wishing to disclose information related to catastrophic risk. S.B. 53, 2025–2026 Leg., Reg. Sess. (Cal. 2025), sec 1107.1.(a). For an example of the type of protocol a company might put in place to protect whistleblowers, see OpenAI, *OpenAI Raising Concerns Policy* (Oct. 2024), https://cdn.openai.com/policies/raising-concerns-policy-blog-copy-202410.pdf.

93 Some best security practices standards to build on include ISO/IEC 27001, NIST 800-53 and SOC 2. ISO/IEC 27001, *Information technology — Security techniques — Information security management systems — Requirements* (2022); Joint Task Force, Nat'l Inst. of Standards & Tech., *Security and Privacy Controls for Information Systems and Organizations* (NIST Spec. Publ. 800-53 rev. 5) (2020), https://doi.org/10.6028/NIST.SP.800-53r5; Am. Inst. of Certified Pub. Accts., *2017 Trust Services Criteria for Security, Availability, Processing Integrity, Confidentiality, and Privacy* (2022). Also see Sella Nevo et al., *Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models* (2024), https://www.rand.org/pubs/research_reports/RRA2849-1.html (to mitigate AI risks in an international context, security level 3 (SL3) and security level 5 (SL5) are particularly worth considering. SL3 is the level needed to prevent most non-state actors from stealing the weights, including most terrorist groups. To prevent world-class state actors top priority operations from stealing the weights - which is key to non-proliferation in an international agreement context - SL5 needs to be reached).

94 Also see Jonas Schuett et al., *Towards best practices in AGI safety and governance: A survey of expert opinion* (May 11, 2023) (unpublished manuscript), https://arxiv.org/abs/2305.07153.

95 For a patchwork of fragments, see: Nat'l Inst. of Standards & Tech., *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile* (NIST.AI.600-1) (2024), https://doi.org/10.6028/NIST.AI.600-1 (suggests to "engage in threat modelling to anticipate potential risks from generative AI systems"); Sella Nevo et al., *Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models* (2024), https://www.rand.org/pubs/research_reports/RRA2849-1.html (which provides a threat model for model weight theft, identifying key threat actors and attack vectors); OpenAI, *Building an early warning system for LLM-aided biological threat creation* (Jan. 31, 2024), https://openai.com/index/building-an-early-warning-system-for-llm-aided-biological-threat-creation/ (which delineates how general-purpose AI capabilities could enable biological threat creation, breaking down one of the two pathways into five distinct stages); OpenAI, *Disrupting malicious uses of AI by state-affiliated threat actors* (Feb. 14, 2024), https://openai.com/index/disrupting-malicious-uses-of-ai-by-state-affiliated-threat-actors/ (identifying existing cyber offence actors who use GPT4 in their activities). Current scenario building efforts predominantly focus on parts of threat scenarios involving malicious actors, leaving gaps in the understanding of risks arising from accidents.

96 Int'l Org. for Standardization & Int'l Electrotechnical Comm'n, *Safety aspects — Guidelines for their inclusion in standards* (ISO/IEC Guide 51:2014) (2014). This standard aligns with the EU's Code of Practice, which calls for analysis of systemic risks in light of "the probability and severity of harm". EUROPEAN COMM'N, *The General-Purpose AI Code of Practice* (July 10, 2025), https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai, 3.4

97 See Int'l Org. for Standardization & Int'l Electrotechnical Comm'n, *Safety aspects — Guidelines for their inclusion in standards* (ISO/IEC Guide 51:2014) (2014).

98 See Int'l Org. for Standardization, *Medical devices — Guidance on the application of ISO 14971*(ISO/TR 24971:2020) (2020) (for guidance on risk estimation); Int'l Org. for Standardization & Int'l Electrotechnical Comm'n, *Guide to the development and inclusion of aspects of safety in International Standards for medical devices* (ISO/IEC Guide 63:2019) (2019) (includes a diagram in Figure 1 that shows the causal chain from hazard to harm); Int'l Electrotechnical Comm'n, *Risk management — Risk assessment techniques* (IEC 31010:2019) (2019) (describes risk assessment techniques).

99 See Siméon Campos et al., *A Frontier AI Risk Management Framework: Bridging the Gap Between Current AI Practices and Established Risk Management*, arXiv:2502.06656 (Feb. 10, 2025), https://arxiv.org/abs/2502.06656, 4-12.

100 See Nathaniel Li et al., *The WMDP Benchmark: Measuring and Reducing Malicious Use With Unlearning* (Mar. 5, 2024) (unpublished manuscript), https://arxiv.org/abs/2403.03218; but consider Dep't for Sci., Innovation & Tech. & AI Safety Inst., *International AI Safety Report 2025* (2025), 186 (indicating that "these proxy measures often do not reliably reflect the true risk in context").

101 See, for example, Mary Phuong et al., *Evaluating Frontier Models for Dangerous Capabilities* (Mar. 20, 2024) (unpublished manuscript), https://arxiv.org/abs/2403.13793; Andy K. Zhang et al., *Cybench: A Framework for Evaluating Cybersecurity Capabilities and Risks of Language Models*(Aug. 15, 2024) (unpublished manuscript), https://arxiv.org/abs/2408.08926.

102 See, for example, Mary Phuong et al., *Evaluating Frontier Models for Dangerous Capabilities* (Mar. 20, 2024) (unpublished manuscript), https://arxiv.org/abs/2403.13793; OpenAI, *OpenAI o3 and o4-mini System Card* (Apr. 16, 2025), https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf, 14.

103 See, for example, Christopher A. Mouton et al., *The Operational Risks of AI in Large-Scale Biological Attacks: Results of a Red-Team Study* (2024), https://www.rand.org/pubs/research_reports/RRA2977-2.html; OpenAI, *OpenAI o3 and o4-mini System Card* (Apr. 16, 2025), https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf, 28.

104 Irene Solaiman et al., *Release Strategies and the Social Impacts of Language Models* (Aug. 24, 2019) (unpublished manuscript), https://arxiv.org/abs/1908.09203; Toby Shevlane et al., *Structured access: an emerging paradigm for safe AI deployment* (Jan. 13, 2022) (unpublished manuscript), https://arxiv.org/abs/2201.05159.

105 See Toby Shevlane et al., *Structured access: an emerging paradigm for safe AI deployment* (Jan. 13, 2022) (unpublished manuscript), https://arxiv.org/abs/2201.05159. This approach, however, by default comes with significant costs to scientific progress on the most advanced systems and reducing power concentration that should be weighted with the cost. Elizabeth Seger et al., *Open-Sourcing Highly Capable Foundation Models: An evaluation of risks, benefits, and alternative methods for pursuing open-source objectives* (Ctr. for the Governance of AI 2023), https://cdn.governance.ai/Open-Sourcing_Highly_Capable_Foundation_Models_2023_GovAI.pdf, 17.

106 See, for example, Anthropic, *Anthropic's Responsible Scaling Policy* (Sep. 19, 2023), https://www.anthropic.com/news/anthropics-responsible-scaling-policy.

107 Joe O'Brien, Shaun Ee & Zoe Williams, *Deployment corrections: An incident response framework for frontier AI models*(Inst. for AI Pol'y & Strategy 2023), https://static1.squarespace.com/static/64edf8e7f2b10d716b5ba0e1/t/651c397fc04af033499df9f8/1696348544356/Deployment+corrections_+an+incident+response+framework+for+frontier+AI+models.pdf.

108 *International Covenant on Economic, Social and Cultural Rights*, Dec. 16, 1966, 993 U.N.T.S. 3, arts 15(1)(b), (2), (3), and (4).

109 *International Covenant on Economic, Social and Cultural Rights*, Dec. 16, 1966, 993 U.N.T.S. 3, art 1(1). See also Ben Saul, David Kinley & Jacqueline Mowbray, *The International Covenant on Economic, Social and Cultural Rights: Commentary, Cases, and Materials* (2014), 67-69.

110 *International Covenant on Civil and Political Rights*, Dec. 16, 1966, 999 U.N.T.S. 171, art 25(1) . Also see U.N. Hum. Rts. Council Res. 39/11, U.N. Doc. A/HRC/RES/39/11 (Oct. 11, 2018); U.N. High Comm'r for Hum. Rts., *Guidelines for States on the effective implementation of the right to participate in public affairs*, U.N. Doc. A/HRC/39/28 (July 11, 2018); U.N. Hum.Rts. Council Res. 33/22, U.N. Doc. A/HRC/RES/33/22 (Oct. 13, 2016).

111 G.A. Res. 41/128, U.N. Doc. A/RES/41/128 (Feb. 23, 1987), para 1 ("The right to development is an inalienable human right by virtue of which every human person and all peoples are entitled to participate in, contribute to, and enjoy economic, social, cultural, and political development, in which all human rights and fundamental freedoms can be fully realized.")

112 Maria Lee, 'The Legal Institutionalization of Public Participation in the EU Governance of Technology' in Roger Brownsword, Eloise Scotford, and Karen Yeung (eds), *The Oxford Handbook of Law, Regulation, and Technology* (OUP 2017), 622.

113 Lee and Abbot, 'The Usual Suspects', 82 referencing *inter alia* Rober Summers, 'Evaluating and Improving Legal Processes - A Plea for Process Values' (1974) 60 Cornell L Rev 1; George R Pring and Susan Y Noé, 'The Emerging International Law of Public Participation Affecting Global Mining, Energy and Resources Development' in Donald Zillman and others (eds), *Human Rights in Natural Resource Development: Public Participation in the Sustainable Development of Mining and Energy Resources* (OUP 2002); OECD, *Framework for Anticipatory Governance of Emerging Technologies*, 11, 16-17, 23-26; See for example 'Guidelines for States on the Effective Implementation of the Right to Participate in Public Affairs', 4; Kathryn S Quick and John M Bryson, 'Public participation' in Christopher K Ansell and Jacob Torfing (eds), *Handbook on Theories of Governance* (Edward Elgar 2016); and Henk Addink, *Good Governance: Concept and Context* (OUP 2019), 129-140; https://icml.cc/virtual/2025/poster/40100#

114 Dep't for Sci., Innovation & Tech. & AI Safety Inst., *International AI Safety Report 2026* (2026); U.N., *Independent International Scientific Panel on AI*, https://www.un.org/independent-international-scientific-panel-ai/.

115 Boaz Miller, *When Is Consensus Knowledge Based? Distinguishing Shared Knowledge from Mere Agreement*, 190 SYNTHESE 1293 (2013).

116 For proposals of a licensing approach, see Anthony Aguirre, *Keep the Future Human* https://keepthefuturehuman.ai/wp-content/uploads/2025/03/Keep_the_Future_Human__AnthonyAguirre__5March2025.pdf, 44; Narrow Path, *Phase 0*, in A Narrow Path (Oct. 2, 2024), https://www.narrowpath.co/phase-0; https://arxiv.org/pdf/2503.18956; more generally, see Matthijs Maas and José Jaime Villalobos, *International AI institutions: A literature review of models, examples, and proposals*, Inst. for L. & AI (Sept. 2023), https://law-ai.org/international-ai-institutions/, 20-25.

117 See Robert Trager et al., *International Governance of Civilian AI: A Jurisdictional Certification Approach*, LawAI Working Paper No. 3-2023 (Aug. 31, 2023), http://dx.doi.org/10.2139/ssrn.4579899; Stephen Lerena, Global Governance of High-Risk AI (working paper).

118 Although it is also worth noting that the field of AI evaluations is still nascent and there are important questions that still need to be addressed as the science of evaluations continues to develop. See Aviv Ovadya et al., *Position: Democratic AI is Possible. The Democracy Levels Framework Shows How It Might Work*, Presentation at the International Conference on Machine Learning (July 15, 2025), https://icml.cc/virtual/2025/poster/40100.

119 See Renan Araujo, *Understanding the First Wave of AI Safety Institutes: Characteristics, Functions, and Challenges*, INST. FOR AI POL'Y & STRATEGY (Oct. 7, 2024), https://www.iaps.ai/research/understanding-aisis.

120 See Robert Trager et al., *International Governance of Civilian AI: A Jurisdictional Certification Approach*, LawAI Working Paper No. 3-2023 (Aug. 31, 2023), http://dx.doi.org/10.2139/ssrn.4579899, 23.

121 The public nature of these reports under article 7 obeys to the right to information, derived from article 19 of the International Covenant on Civil and Political Rights and other international human rights conventions. It embraces a right of access to information held by public bodies, meaning all branches of the State and other public or governmental authorities, at whatever level. To give effect to the right of access to information, States should proactively put in the public domain government information of "public interest" and make every effort to ensure easy, prompt, effective and practical access to such information.# In particular, it is crucial for personal decision-making, participation in democratic decision-making processes, and for the assessment of public authorities' performance, to hold them accountable and prevent corruption. States are further obligated to monitor human rights developments within their jurisdiction and to adequately inform their citizens about relevant human rights risks. The European Court of Human Rights has emphasised the importance of public access to information "which would enable members of the public to assess the danger to which they are exposed", although the Court has not explicitly linked this positive obligation to its equivalent right to information provision. Instead, this obligation exists as a function of guaranteeing individual rights, such as the right to life. Article 19 ICCPR, Article 10 of the Convention for the Protection of Human Rights and Fundamental Freedoms (European Convention on Human Rights, as amended) (ETS No 5) (adopted 4 November 1950); and Article 13 of the American Convention on Human Rights 'Pact of San José, Costa Rica' (adopted 22 November 1969, entry into force 18 July 1978) 1144 UNTS 123; HRC General Comment No. 34, Article 19 (Freedoms of opinion and expression) (12 September 2011) UN Doc CCPR/C/GC/34, [18]-[19]. See also UNCHR Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression (18 January 2000) UN Doc E/CN.4/2000/63, [42]-[44]; UNCHR Res 2000/38 The Right to Freedom of Opinion and Expression (20 April 2000), [10(a)]; Maria Monnheimer, *Due Diligence Obligations in International Human Rights Law* (CUP 2021), 233-236; Diane Desierto, 'Austerity Measures and International Economic, Social, and Cultural Rights' in Evan Criddle (ed), *Human Rights in Emergencies* (CUP 2016), 246; Dinah Shelton and Ariel Gould, 'Positive and Negative Obligations' in Diana Shelton (ed), The Oxford Handbook of International Human Rights Law (OUP 2015), 572; *Taškin and Others v Turkey* App no 46117/99 (ECtHR, 10 November 2004), [119]; *Budayeva and Others v Russia* App no 15339/02 (ECtHR, 20 March 2008), [132]; *Kolyadenko and Others v Russia* App nos 17423/05, 20534/05, 20678/05, 23263/05, 24283/05 and 35673/05 (ECtHR, 28 February 2012), [157]-[161], *Brincat and Others v Malta* App no 60908/11(ECtHR, 24 July 2014), [101]-[102]. See also Linos-Alexander Sicilianos, 'Preventing Violations of the Right to Life: Positive Obligations under Article 2 of the ECHR' (2014) 3(2) Cyprus Hum Rts L R 117, 127.

122 Dep't for Sci., Innovation & Tech. & AI Safety Inst., *International AI Safety Report 2026* (2026), 72-87.

123 See Elika Somani et al., *Strengthening Emergency Preparedness and Response for AI Loss of Control Incidents* (RAND Corp., 2025), https://www.rand.org/pubs/research_reports/RRA3847-1.html.

124 This definition borrows from others used by national and international agencies with emergency response functions. See U.N. Off. for Disaster Risk Reduction, *Sendai Framework Terminology on Disaster Risk Reduction*, https://www.undrr.org/drr-glossary/terminology; World Health Org., *Emergency Response Framework (ERF), Edition 2.1* (2024), https://www.who.int/publications/i/item/9789240058064; Int'l Atomic Energy Agency, *IAEA Nuclear Safety and Security Glossary: 2022 (Interim) Edition* (2022), https://www.iaea.org/resources/publications/iaea-nuclear-safety-and-security-glossary; Fed. Emergency Mgmt. Agency, *Acronyms, Abbreviations, & Terms: A Capability Assurance Job Aid*, https://www.fema.gov/pdf/plan/glo.pdf.

125 Elika Somani et al., *Strengthening Emergency Preparedness and Response for AI Loss of Control Incidents* (RAND Corp., 2025), https://www.rand.org/pubs/research_reports/RRA3847-1.html, 22.

126 For examples of similar incident trackers and databases, see MIT, *MIT AI Incident Tracker* (2025), https://airisk.mit.edu/ai-incident-tracker; OECD, *OECD AIM: AI Incidents and Hazards Monitor*, OECD.AI (2025), https://oecd.ai/en/incidents; Ctr. for Sec. & Emerging Tech., *AI Incident Database*, https://incidentdatabase.ai/; *AI Risk Explorer*, https://www.airiskexplorer.com/.

127 José Jaime Villalobos, Matthijs Maas and Christoph WInter, *International Catastrophe Law: from Nuclear Weapons to Advanced AI* (Cambridge University Press, forthcoming 2026), ch 5.2.3.3.

128 For more on coordination during times of emergency, see Laura G. Militello et al., *Large-Scale Coordination in Emergency Response*, 49 Proc. Hum. Factors & Ergonomics Soc'y Ann. Meeting 534 (2005).

129 On the right to access to information in times of crisis, see commentary under article 7, paragraph 12.

130 See *Universal Periodic Review*, U.N. Off. of the High Comm'r for Hum. Rts., https://www.ohchr.org/en/hr-bodies/upr/upr-home; *Development co-operation peer reviews and learning*, Org. for Econ. Co-op. & Dev., https://www.oecd.org/en/topics/sub-issues/development-co-operation-peer-reviews-and-learning.html; *Implementation Review Mechanism*, U.N. Off. on Drugs & Crime, https://www.unodc.org/corruption/en/uncac/implementation-review-mechanism.html.

131 U.N. Hum. Rts. Council, *Universal Periodic Review Facts & Figures 2024* (Dec. 2024), https://www.ohchr.org/sites/default/files/documents/hrbodies/upr/UPR-Fact-and-Figures-2024.pdf (based on figures from the end of 2024)

132 U.N. Dev. Programme, U.N. Off. of the High Comm'r for Hum. Rts. & U.N. Dev. Coordination Off., *UN good practices: How the universal periodic review process supports sustainable development* (Feb. 2022), https://www.ohchr.org/sites/default/files/2022-02/UPR_good_practices_2022.pdf.

133 *Development co-operation peer reviews and learning*, Org. for Econ. Co-op. & Dev., https://www.oecd.org/en/topics/sub-issues/development-co-operation-peer-reviews-and-learning.html.

134 U.N. Info. Serv. Vienna, *Real change generated by the Implementation Review Mechanism for the anti-corruption Convention*, UNIS/CP/1089 (Dec. 19, 2019), https://unis.unvienna.org/unis/en/pressrels/2019/uniscp1089.html.

135 U.N. Off. for Disaster Risk Reduction, *Sendai Framework for Disaster Risk Reduction 2015–2030* (2015), https://www.undrr.org/media/16176/download?startDownload=20251105, Priority 4.

136 Int'l Law Comm'n, *Draft Articles on Responsibility of States for Internationally Wrongful Acts, with commentaries*, U.N. Doc. A/56/10 (2001), *reprinted in* [2001] 2 Y.B. Int'l L. Comm'n (pt. 2) 26, https://legal.un.org/ilc/texts/instruments/english/commentaries/9_6_2001.pdf, art 2 and commentary.

137 José Jaime Villalobos, Matthijs Maas and Christoph WInter, *International Catastrophe Law: from Nuclear Weapons to Advanced AI* (Cambridge University Press, forthcoming 2026), ch 4.

138 On the right to international solidarity, see U.N. Off. of the High Comm'r for Hum. Rts., *Draft declaration on the right to international solidarity*, https://www.ohchr.org/en/special-procedures/ie-international-solidarity/draft-declaration-right-international-solidarity(last visited Nov. 5, 2025).

139 Rüdiger Wolfrum, *International Law of Cooperation*, Max Planck Encyclopedia of Pub. Int'l L. (Mar. 2011), https://opil.ouplaw.com/display/10.1093/law:epil/9780199231690/law-9780199231690-e1427, para 2; *Declaration on Principles of International Law concerning Friendly Relations and Co-operation among States in accordance with the Charter of the United Nations*, G.A. Res. 2625 (XXV), U.N. GAOR, 25th Sess., Supp. No. 18, U.N. Doc. A/8028, at 121 (Oct. 24, 1970), art 4.

140 An example of this type of cooperation is the International Network of AI Safety Institutes' collaboration on agentic evaluations. AI Sec. Inst., *International joint testing exercise: Agentic testing* (Jul. 17, 2025), https://www.aisi.gov.uk/work/international-joint-testing-exercise-agentic-testing.

141 Sumaya Nur Adan et al., *AI Benefit-Sharing Framework: Balancing Access and Safety*, Oxford Martin AIGI (Dec. 1, 2025), https://aigi.ox.ac.uk/publications/ai-benefit-sharing-framework-balancing-access-and-safety/, 67-68.

142 Sumaya Nur Adan et al., *AI Benefit-Sharing Framework: Balancing Access and Safety*, Oxford Martin AIGI (Dec. 1, 2025), https://aigi.ox.ac.uk/publications/ai-benefit-sharing-framework-balancing-access-and-safety/, 66-67.

143 See Hannah Chafetz, Andrew J. Zahuranec & Stefaan G. Verhulst, *A Blueprint to Unlock New Data Commons for AI* (Open Data Policy Lab, 2025), https://incubator.opendatapolicylab.org/files/data-commons-for-ai-blueprint.pdf.

144 See Anthony Aguirre, *Keep the Future Human* https://keepthefuturehuman.ai/wp-content/uploads/2025/03/Keep_the_Future_Human__AnthonyAguirre__5March2025.pdf. For examples of AI for good, see International Telecommunications Union and Oxford Martin AI Governance Initiative, *The Annual AI Governance Report 2025: Steering the Future of AI* (Oct 7, 2025), https://oms-www.files.svdcdn.com/production/images/ai-standards-for-global-impact1.pdf?dm=1759309533 (specifically, see use cases in Part II).

145 See Alan Tang et al., *AI for Public Good: From Vision to Action and Innovation*, U.N. Dev. Programme: Sing. Glob. Ctr. (Jan. 27, 2025), https://www.undp.org/policy-centre/singapore/blog/ai-public-good-vision-action-and-innovation.

146 See Int'l Dialogues on AI Safety, *IDAIS-Venice, 2024*, https://idais.ai/dialogue/idais-venice/ (last visited Nov. 5, 2025); Kayla Blomquist et al., *Examining AI Safety as a Global Public Good: Implications, Challenges, and Research Priorities*, Carnegie Endowment for Int'l Peace (Mar. 11, 2025), https://carnegieendowment.org/research/2025/03/examining-ai-safety-as-a-global-public-good-implications-challenges-and-research-priorities?lang=en; Ben Bucknall et al., *In Which Areas of Technical AI Safety Could Geopolitical Rivals Cooperate?*, arXiv:2504.12914 (2025), https://arxiv.org/abs/2504.12914.

147 See, for example, U.N. Comm. of Experts on Big Data & Data Sci. for Off. Stat., Task Team on Privacy-Enhancing Techs., *Privacy-Enhancing Technologies*, https://unstats.un.org/bigdata/task-teams/privacy/index.cshtml.

148 Matthew Sargent et al., *Assessing the Value of a CERN-Like Multinational Research Organization for AI*, PE-A4024-1 (RAND Corp., 2025), https://www.rand.org/pubs/perspectives/PEA4024-1.html, 5-11.

149 See, for example,Yoshua Bengio et al., *Superintelligent Agents Pose Catastrophic Risks: Can Scientist AI Offer a Safer Path?*, arXiv:2502.15657 (2025), https://arxiv.org/abs/2502.15657.

150 A Narrow Path, *Phase 0*, in A Narrow Path (Oct. 2, 2024), https://www.narrowpath.co/phase-0; https://arxiv.org/pdf/2503.18956; Jason Hausenloy, Andrea Miotti & Claire Dennis, *Multinational AGI Consortium (MAGIC): A Proposal for International Coordination on AI*, arXiv:2310.09217 (Oct. 13, 2023), https://arxiv.org/abs/2310.09217; Duncan Cass-Beggs, Matthew da Mota & Abhiram Reddy, *A Joint International AI Lab: Design Considerations*, CIGI Papers No. 334 (Oct. 2025), https://www.cigionline.org/static/documents/no.334.pdf; Matthijs Maas and José Jaime Villalobos, *International AI institutions: A literature review of models, examples, and proposals*, Inst. for L. & AI (Sept. 2023), https://law-ai.org/international-ai-institutions/, 27-30.

151 See, for example, Alex Petropoulos et al., *Building CERN for AI: An institutional blueprint*, Centre for Future Generations (Jan. 30, 2025), https://cfg.eu/building-cern-for-ai/.

152 See, for example, Cecilia Rikap et al., *Reclaiming digital sovereignty: A roadmap to build a digital stack for people and the planet*, UCL Inst. for Innovation & Pub. Purpose (2024), https://www.ucl.ac.uk/bartlett/publications/2024/dec/reclaiming-digital-sovereignty; Johannes et al., *INTELLECT-1 Release: The First Globally Trained 10B Parameter Model*, Prime Intellect Blog (Nov. 29, 2024), https://www.primeintellect.ai/blog/intellect-1-release.

153 See Kiran Tomlinson et al., *Working with AI: Measuring the Applicability of Generative AI to Occupations*, arXiv:2507.07935 (July 10, 2025), https://arxiv.org/abs/2507.07935; https://www.convergenceanalysis.org/threshold-2030/part-1-worldbuilding#significant-unemployment.

154 For a related analysis of the erosion of mid-20th century social security institutions in the face of seismic economic shifts, see Samuel Moyn, *Not Enough: Human Rights in an Unequal World* (2019). For an exploration of social dislocation in the face of emerging technologies, see Karl Polanyi, *The Great Transformation: The Political and Economic Origins of Our Time* (Penguin Classics 2024); and for parallels with AI adoption, see Daron Acemoglu & Simon Johnson, *Learning From Ricardo and Thompson: Machinery and Labor in the Early Industrial Revolution and in the Age of Artificial Intelligence*, 16 Ann. Rev. Econ. 597 (2024), https://doi.org/10.1146/annurev-economics-091823-025129.

155 See Ramishah Maruf, *Takeaways from Anthropic CEO Dario Amodei's CNN interview*, CNN (May 29, 2025), https://kvia.com/news/business-technology/cnn-business-consumer/2025/05/29/why-this-leading-ai-ceo-is-warning-the-tech-could-cause-mass-unemployment-2/ (for an explanation of this phenomenon by Dario Amodei, the CEO of Anthropic, one of the leading AI corporations in the United States).

156 See Anton Korinek, Martin Schindler & Joseph Stiglitz, *Technological Progress, Artificial Intelligence, and Inclusive Growth*, IMF Working Paper No. 2021/166 (June 11, 2021), https://www.imf.org/en/Publications/WP/Issues/2021/06/11/Technological-Progress-Artificial-Intelligence-and-Inclusive-Growth-460695.

157 See David Rotman, *How to Solve AI's Inequality Problem*, MIT Tech. Rev. (Apr. 19, 2022), https://www.technologyreview.com/2022/04/19/1049378/ai-inequality-problem/; S. Bushwick, *Unregulated AI Will Worsen Inequality, Warns Nobel-Winning Economist Joseph Stiglitz*, Sci. Am. (Aug. 1, 2023), https://www.scientificamerican.com/article/unregulated-ai-will-worsen-inequality-warns-nobel-winning-economist-joseph-stiglitz/**;** Eugenio M. Cerutti et al., The Global Impact of AI: Mind the Gap (Int'l Monetary Fund, Working Paper No. 25/76, 2025), available at https://www.imf.org/en/Publications/WP/Issues/2025/04/11/The-Global-Impact-of-AI-Mind-the-Gap-566129.

158 See Chinasa T. Okolo, *AI in the Global South: Opportunities and challenges towards more inclusive governance*, Brookings (Nov. 1, 2023), https://www.brookings.edu/articles/ai-in-the-global-south-opportunities-and-challenges-towards-more-inclusive-governance/.

159 Anna Yelizarova, *The Missing Institution: A Global Dividend System for the Age of AI*, AGI Social Contract (May 28, 2025), https://www.agisocialcontract.org/anthology/windfall.

160 See Andreas Schmidt & Daan Juijn, *Economic inequality and the long-term future*, 23 POL. PHIL. & ECON. 67 (2024), https://doi.org/10.1177/1470594X231178502.

161 See Windfall Trust, *Windfall Policy Atlas* (2026) https://windfalltrust.org/policy-atlas; Sumaya Nur Adan, Joanna Wiaterek et al., *AI Benefit-Sharing Framework: Balancing Access and Safety*, Oxford Martin AIGI (Dec. 1, 2025), https://aigi.ox.ac.uk/publications/ai-benefit-sharing-framework-balancing-access-and-safety/; Claire Dennis et al., *Options and Motivations for International AI Benefit Sharing*, Centre for the Governance of AI (Feb. 6, 2025), https://www.governance.ai/research-paper/options-and-motivations-for-international-ai-benefit-sharing; Emma Casey et al., *Designing an AI Bond for Growth and Shared Prosperity in the UK*, Centre for British Progress (July 2, 2024), https://britishprogress.org/uk-day-one/designing-an-ai-bond-for-growth-and-shared-prosper.

162 For a proposal on the distribution of these gains or "windfall", see Cullen O'Keefe et al., *The Windfall Clause: Distributing the Benefits of AI for the Common Good* (Centre for the Governance of AI & Future of Humanity Inst. 2020), https://cdn.governance.ai/Windfall-Clause-Report.pdf; for some estimates of the cost of a global dividend under different scenarios of economic transformation due to AI, see Anna Yelizarova, *The Missing Institution: A Global Dividend System for the Age of Transformative AI* (The Digitalist Papers 2025), https://www.digitalistpapers.com/vol2/yelizarova.

163 Cullen O'Keefe et al., *The Windfall Clause: Distributing the Benefits of AI for the Common Good* (Centre for the Governance of AI & Future of Humanity Inst. 2020), https://cdn.governance.ai/Windfall-Clause-Report.pdf, 40.

164 Cullen O'Keefe et al., *The Windfall Clause: Distributing the Benefits of AI for the Common Good* (Centre for the Governance of AI & Future of Humanity Inst. 2020), https://cdn.governance.ai/Windfall-Clause-Report.pdf, 43.

165 Int'l Law Comm'n, *Draft Articles on Responsibility of States for Internationally Wrongful Acts, with commentaries*, U.N. Doc. A/56/10 (2001), *reprinted in* [2001] 2 Y.B. Int'l L. Comm'n (pt. 2) 26, https://legal.un.org/ilc/texts/instruments/english/commentaries/9_6_2001.pdf, 40-41.

166 The Biological Weapons Convention obliges states to pass domestic laws prohibiting the development, stockpiling or use of these weapons. Convention on the Prohibition of the Development, Production and Stockpiling of Bacteriological (Biological) and Toxin Weapons and on Their Destruction, Apr. 10, 1972, 26 U.S.T. 583, 1015 U.N.T.S. 163, art 4. This mandate has led to a wave of national legislation. The United States implements this obligation through 18 U.S.C 175, which penalizes the possession or dissemination of biological agents as weapons. For an enforcement of this rule, see Jason Pate & Jacqueline E. Miller, *Minnesota Patriots Council*, *in* ENCYCLOPEDIA OF BIOTERRORISM DEFENSE (Richard F. Pilch & Raymond A. Zilinskas eds., 2011).
In the UK, the Biological Weapons Act 1974 criminalizes similar conduct. Biological Weapons Act 1974, c. 6 (U.K.). Similarly, the Chemical Weapons Convention mandates domestic legal sanctions against chemical weapon research and deployment. Convention on the Prohibition of the Development, Production, Stockpiling and Use of Chemical Weapons and on Their Destruction, Jan. 13, 1993, 1974 U.N.T.S. 45, art 7. Also see Thomas Brown, *Judicial Enforcement of BWC and CWC implementing legislation*, VERTIC Brief No. 34 (Feb. 2022), https://www.vertic.org/wp-content/uploads/2022/02/VERTIC-Brief-34_TB_final.pdf.

167 For example, under the EU AI Act, violations may incur administrative fines of up to 7% of global annual turnover. Regulation (EU) 2024/1689, 2024 O.J. (L 2024/1689) 1, art 99.3.
Unlike criminal law, such fines do not require proof of culpability or mens rea, making enforcement swifter and more certain. Martin Braun et al, *Prohibited AI Practices: A Deep Dive Into Article 5 of the European Union's AI Act*, WilmerHale Privacy and Cybersecurity Law (Apr. 8, 2024), https://www.wilmerhale.com/en/insights/blogs/wilmerhale-privacy-and-cybersecurity-law/20240408-prohibited-ai-practices-a-deep-dive-into-article-5-of-the-european-unions-ai-act.

168 K. Mamak, *AGI crimes? The role of criminal law in mitigating existential risks posed by artificial general intelligence*, 40 AI & SOC'Y 2691 (2025), https://doi.org/10.1007/s00146-024-02036-5.

169 Europol Innovation Lab, *Facing Reality? Law Enforcement and the Challenge of Deepfakes* (2022), https://www.europol.europa.eu/cms/sites/default/files/documents/Europol_Innovation_Lab_Facing_Reality_Law_Enforcement_And_The_Challenge_Of_Deepfakes.pdf, 14.

170 A. Mahmud, *Application and Criminalization of Artificial Intelligence in the Digital Society: Security Threats and the Regulatory Challenges*, 18 J. APPIED SEC. RES. 1 (2021), https://doi.org/10.1080/19361610.2021.1947113; Matthias Schulze, *The encryption debate: 40 years, the same arguments*, TRESORIT BLOG (Oct. 3, 2016), https://cms.tresorit.com/blog/encryption-debate.

171 Athina Sachoulidou, *AI Systems and Criminal Liability: A Call for Action*, 11 OSLO L. REV. 1 (2024), https://doi.org/10.18261/olr.11.1.3.

172 See, for example, Cartagena Protocol on Biosafety to the Convention on Biological Diversity, Jan. 29, 2000, 2226 U.N.T.S. 208, 39 I.L.M. 1027, art 19; Basel Convention on the Control of Transboundary Movements of Hazardous Wastes and Their Disposal, Mar. 22, 1989, 1673 U.N.T.S. 57, 28 I.L.M. 649, art 5; Arms Trade Treaty, Apr. 2, 2013, 3013 U.N.T.S. 269, 52 I.L.M. 988, art 5.

173 Renan Araujo, *Understanding the First Wave of AI Safety Institutes: Characteristics, Functions, and Challenges*, INST. FOR AI POL'Y & STRATEGY. (Oct. 7, 2024), https://www.iaps.ai/research/understanding-aisis.

174 Intn'l Network of AI Safety Insts, *Mission Statement* (Nov. 20, 2024), https://www.nist.gov/system/files/documents/2024/11/20/Mission%20Statement%20-%20International%20Network%20of%20AISIs.pdf.

175 Marta Ziosi et al., Oxford Martin AI Governance Initiative, *AISIs' Roles in Domestic and International Governance* (July 12, 2024), https://www.oxfordmartin.ox.ac.uk/publications/aisis-roles-in-domestic-and-international-governance; Frank Ryan, Niki Iliadis & George Gor, *Weakening a Safety Net: Key Considerations for How the AI Safety Institute Network Can Advance Multilateral Collaboration*, The Future Society (2024), https://thefuturesociety.org/wp-content/uploads/2024/11/Weaving-a-Safety-Net_AISI-Collaboration_November-2024.pdf, 4. For other proposals on membership expansion, see Gregory C. Allen, *A Safety Network: The Future of AI Safety Institute Collaboration,* Ctr. for Strategic & Int'l Stud. (Oct. 30, 2024), https://csis-website-prod.s3.amazonaws.com/s3fs-public/2024-10/241030_Allen_Safety_Network.pdf., 16; Sumaya Nur Adan et al., *Key Questions for the International Network of AI Safety Institutes* (Inst. for AI Pol'y & Strategy, Nov. 9, 2024), https://static1.squarespace.com/static/64edf8e7f2b10d716b5ba0e1/t/672f4baae2c5d504ce1cacca/1731152811021/IAPS+-+Key+questions+for+the+International+Network+of+AI+Safety+Institutes.pdf, 10.

176 See a similar proposal in Frank Ryan, Niki Iliadis & George Gor, *Weakening a Safety Net: Key Considerations for How the AI Safety Institute Network Can Advance Multilateral Collaboration* (The Future Society, 2024), https://thefuturesociety.org/wp-content/uploads/2024/11/Weaving-a-Safety-Net_AISI-Collaboration_November-2024.pdf, 6.

177 See a similar proposal in Robert Trager et al., *International Governance of Civilian AI: A Jurisdictional Certification Approach*, LawAI Working Paper No. 3-2023 (Aug. 31, 2023), http://dx.doi.org/10.2139/ssrn.4579899,

178 Matthijs Maas and José Jaime Villalobos, *International AI institutions: A literature review of models, examples, and proposals*, Inst. for L. & AI (Sept. 2023), https://law-ai.org/international-ai-institutions/, 14-18.

179 See, for example, Convention on the Prohibition of the Dev., Prod., Stockpiling & Use of Chem. Weapons & on Their Destruction, Jan. 13, 1993, 1975 U.N.T.S. 45, 32 I.L.M. 800., art 13(22); Arms Trade Treaty, Apr. 2, 2013, 3013 U.N.T.S. 269, 52 I.L.M. 988, art 17(1); United Nations Framework Convention on Climate Change, May 9, 1992, 1771 U.N.T.S. 107, 31 I.L.M. 849, art 7(4).

180 See Matthijs Maas and José Jaime Villalobos, *International AI institutions: A literature review of models, examples, and proposals*, Inst. for L. & AI (Sept. 2023), https://law-ai.org/international-ai-institutions/, 20-26.

181 See Claire Dennis et al., *Options and Motivations for International AI Benefit Sharing*, Centre for the Governance of AI (Feb. 6, 2025), https://www.governance.ai/research-paper/options-and-motivations-for-international-ai-benefit-sharing, 35.

182 For general rules on amendments of international agreements, see Vienna Convention on the Law of Treaties, May 23, 1969, 1155 U.N.T.S. 331, 8 I.L.M. 679, arts 39-41.

183 See, for example, Convention on the Prohibition of the Development, Production, Stockpiling and Use of Chemical Weapons and on Their Destruction, Jan. 13, 1993, 1974 U.N.T.S. 45, art 15(2); Arms Trade Treaty, Apr. 2, 2013, 3013 U.N.T.S. 269, art 20(2); Rome Statute of the International Criminal Court, July 17, 1998, 2187 U.N.T.S. 90, art 121(3).

184 Vienna Convention on the Law of Treaties, May 23, 1969, 1155 U.N.T.S. 331, 8 I.L.M. 679, art 41.1(a)(ii).

185 See James Devaney, Fact-Finding and Expert Evidence, in The Cambridge Companion to the International Court of Justice 187 (Carlos Espósito & Kate Parlett eds., 2023); Makane Moïse Mbengue, Scientific Fact-finding at the International Court of Justice: An Appraisal in the Aftermath of the Whaling Case, 29 Leiden J. Int'l L. 529 (2016).

186 Established under United Nations Convention on the Law of the Sea, Dec. 10, 1982, 1833 U.N.T.S. 397, art 287.

187 See Helmut Tuerk, 20 Years of the International Tribunal for the Law of the Sea (ITLOS): An Overview, 49 Rev. belge droit int'l 449 (2016); Nilüfer Oral, The Contribution of ITLOS to the Development of International Law for Protection of the Marine Environment and Conservation of Living Resources, in Case-Law and the Development of International Law 180 (Patrícia Galvão Teles & Manuel Almeida Ribeiro eds., 2021); Philippe Gautier, Experts before ITLOS: An Overview of the Tribunal's Practice, 9 J. Int'l Disp. Settle. 433 (2018); Natalie Klein, The Effectiveness of the UNCLOS Dispute Settlement Regime: Reaching for the Stars?, 108 ASIL Proc. 359 (2014).

188 For analysis of the way that various international courts decide requests for provisional measures, see Cameron A. Miles, Provisional Measures before International Courts and Tribunals (2017).

189 Treaty on the Non-Proliferation of Nuclear Weapons, July 1, 1968, 21 U.S.T. 483, 729 U.N.T.S. 161, art VI; Treaty on the Prohibition of Nuclear Weapons, adopted July 7, 2017, U.N. Doc. A/CONF.229/2017/8, art 12.

190 José Jaime Villalobos, Matthijs Maas and Christoph WInter, *International Catastrophe Law: from Nuclear Weapons to Advanced AI* (Cambridge University Press, forthcoming 2026), chs 3, 4.

191 Vienna Convention on the Law of Treaties, May 23, 1969, 1155 U.N.T.S. 331, 8 I.L.M. 679, art 31(3)(c) ("There shall be taken into account, together with the context: … any relevant rules of international law applicable in the relations between the parties.")

192 See Jean d'Aspremont, *The Systemic Integration of International Law by Domestic Courts: Domestic Judges as Architects of the Consistency of the International Legal Order*, *in* The Practice of International and National Courts and the (De-)Fragmentation of International Law 150 (Ole Kristian Fauchald & André Nollkaemper eds., 2012); Int'l L. Comm'n, *Fragmentation of International Law: Difficulties Arising from the Diversification and Expansion of International Law: Report of the Study Group of the International Law Commission, finalized by Martti Koskenniemi*, U.N. Doc. A/CN.4/L.682 (Apr. 13, 2006), Annex A, B(a)(4).

193 Tereza Zoumpalova & Niki Iliadis, *Part 1: What Are Red Lines for AI and Why Are They Important?*, The Future Society (July 3, 2025), https://thefuturesociety.org/airedlines-partone. Also see Members of the Glob. Future Council on the Future of AI, *AI red lines: the opportunities and challenges of setting limits*, WORLD ECON. F. (Mar. 11, 2025), https://www.weforum.org/stories/2025/03/ai-red-lines-uses-behaviours/; Eunseo Dana Choi & Dylan Rogers, *Risk thresholds for frontier AI: Insights from the AI Action Summit*, OECD.AI (Mar. 5, 2025), https://oecd.ai/en/wonk/risk-thresholds-for-frontier-ai-insights-from-the-ai-action-summit.

194 U.N. GAOR, 78th Sess., 3d Comm., U.N. Doc. A/C.3/78/L.19/Rev.1, *Rights of the child* (Nov. 8, 2023), https://documents.un.org/doc/undoc/ltd/n23/342/80/pdf/n2334280.pdf.

195 Regulation (EU) 2024/1689, 2024 O.J. (L 2024/1689) 1, art 5.

196 TAKE IT DOWN Act, Pub. L. No. 119-12 (2025).

197 S. Comm. on Bus. & Com., Bill Analysis, H.B. 149, 89th Leg., R.S. (Tex. 2025), https://capitol.texas.gov/tlodocs/89R/analysis/html/HB00149S.htm.

198 Hum. Rts. Watch, *Stopping Killer Robots: Country Positions on Banning Fully Autonomous Weapons and Retaining Human Control* (Aug. 10, 2020), https://www.hrw.org/report/2020/08/10/stopping-killer-robots/country-positions-banning-fully-autonomous-weapons-and.

199 *International Dialogue on AI Safety (IDAIS): Beijing*, IDAIS, https://idais.ai/dialogue/idais-beijing/.

200 EUROPEAN COMM'N, *The General-Purpose AI Code of Practice* (July 10, 2025), https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai, Annex 1.4.

201 Deepika Raman et al., *Intolerable Risk Threshold Recommendations for Artificial Intelligence*, UC Berkeley Ctr. for Long-Term Cybersecurity (Feb. 2025), https://cltc.berkeley.edu/wp-content/uploads/2025/02/Intolerable-Risk-Threshold-Recommendations-for-Artificial-Intelligence.pdf, 26

202 For examples of this type of prohibition, see Treaty on the Prohibition of Nuclear Weapons, July 7, 2017, 3370 U.N.T.S. (registration No. 56487); Convention on the Prohibition of the Development, Production and Stockpiling of Bacteriological (Biological) and Toxin Weapons and on Their Destruction, Apr. 10, 1972, 26 U.S.T. 583, 1015 U.N.T.S. 163; Convention on the Prohibition of the Development, Production, Stockpiling and Use of Chemical Weapons and on Their Destruction, Jan. 13, 1993, 1974 U.N.T.S. 45.

203 José Jaime Villalobos, Matthijs Maas and Christoph WInter, *International Catastrophe Law: from Nuclear Weapons to Advanced AI* (Cambridge University Press, forthcoming 2026), ch 4.

204 Anthropic, *Activating AI Safety Level 3 protections*, ANTHROPIC (May 22, 2025), https://www.anthropic.com/news/activating-asl3-protections; OpenAI, *Preparing for future AI capabilities in biology*, OPENAI (June 18, 2025), https://openai.com/index/preparing-for-future-ai-capabilities-in-biology/.

205 UK AI Security Institute, *Our Evaluation of Claude Mythos Preview's Cyber Capabilities*, https://www.aisi.gov.uk/blog/our-evaluation-of-claude-mythos-previews-cyber-capabilities (Apr. 13, 2026).

206 See Anthony Aguirre, Control Inversion (Nov. 6, 2025), https://control-inversion.ai/.

207 Deepika Raman et al., *Intolerable Risk Threshold Recommendations for Artificial Intelligence*, UC Berkeley Ctr. for Long-Term Cybersecurity (Feb. 2025), https://cltc.berkeley.edu/wp-content/uploads/2025/02/Intolerable-Risk-Threshold-Recommendations-for-Artificial-Intelligence.pdf, 28.

208 Anthony Aguirre, *Control Inversion* (Nov. 6, 2025), https://control-inversion.ai/; Daniel Kokotajlo & Eli Lifland, *Takeoff Forecast*, AI 2027 (Apr. 2025), https://ai-2027.com/research/takeoff-forecast.

209 Jan Kulveit et al., *Gradual Disempowerment: Systemic Existential Risks from Incremental AI Development*, arXiv:2501.16946 (Jan. 28, 2025), https://arxiv.org/abs/2501.16946.

210 *International Dialogue on AI Safety (IDAIS): Beijing*, IDAIS, https://idais.ai/dialogue/idais-beijing/.

211 Dep't for Sci., Innovation & Tech. & AI Safety Inst., *International AI Safety Report 2025* (2025), 3.2.1; Oscar Delaney, Oliver Guest & Renan Araujo, *Policy Options for Preserving Chain of Thought Monitorability*, Inst. for AI Pol'y & Strategy (Sept. 24, 2024), https://www.iaps.ai/research/policy-options-for-preserving-cot-monitorability?rq=chain.

212 Anthropic, *Responsible Scaling Policy* v3.1, at 8–9 (Apr. 2, 2026), https://www.anthropic.com/responsible-scaling-policy (identifying "high-stakes sabotage opportunities" as a capability threshold distinct from loss of control, defined as AI systems "highly relied upon" with "extensive access to sensitive assets" and "moderate capacity for autonomous, goal-directed operation and subterfuge").

213 Anthropic, *Pilot Sabotage Risk Report* (Oct. 28, 2025), https://alignment.anthropic.com/2025/sabotage-risk-report/ (noting that as AI systems take on research and analysis roles in risk assessment, "analyses of threats from AIs should follow very high evidentiary standards" because "key evidence may be suspect").

214 Alexander Meinke et al., *Frontier Models are Capable of In-Context Scheming*, APOLLO RESEARCH (Dec. 2024), https://arxiv.org/abs/2412.04984; *see also* APOLLO RESEARCH & OPENAI, *Stress Testing Deliberative Alignment for Anti-Scheming Training* (Sept. 2025), https://www.apolloresearch.ai/research/stress-testing-deliberative-alignment-for-anti-scheming-training (documenting "lying, sabotaging of useful work, sandbagging in evaluations, [and] reward hacking" across o3, o4-mini, Gemini 2.5 Pro, Claude 4 Opus, and Grok 4).

215 Regulation (EU) 2024/1689, 2024 O.J. (L 2024/1689) 1, art 5.1. For commentary on the justification for this prohibition under the Act, see *Recitals - AI Act*, AI-ACT-LAW.EU, https://ai-act-law.eu/recital/, recital 29.

216 Deepika Raman et al., *Intolerable Risk Threshold Recommendations for Artificial Intelligence*, UC Berkeley Ctr. for Long-Term Cybersecurity (Feb. 2025), https://cltc.berkeley.edu/wp-content/uploads/2025/02/Intolerable-Risk-Threshold-Recommendations-for-Artificial-Intelligence.pdf, 28; Tom Davidson et al., *AI-Enabled Coups: How a Small Group Could Use AI to Seize Power*, Forethought (Apr. 15, 2025), https://www.forethought.org/research/ai-enabled-coups-how-a-small-group-could-use-ai-to-seize-power; Steven Lee Myers & Stuart A. Thompson, *A.I. Is Starting to Wear Down Democracy*, N.Y. TIMES (June 26, 2025), https://www.nytimes.com/2025/06/26/technology/ai-elections-democracy.html.

217 Treaty on the Reduction and Limitation of Strategic Offensive Arms, U.S.-U.S.S.R., July 31, 1991, T.I.A.S. No. 12,867 [START] (created provisions for baseline data inspections and snap "suspect-site" inspections among a dozen types of on-site checks);Treaty on the Non-Proliferation of Nuclear Weapons, opened for signature July 1, 1968, 21 U.S.T. 483, 729 U.N.T.S. 161 (mandates that non-nuclear weapon states accept IAEA safeguards on their civilian nuclear programs, creating a standing international inspection system); Convention on the Prohibition of the Development, Production, Stockpiling and Use of Chemical Weapons and on Their Destruction, Jan. 13, 1993, 1974 U.N.T.S. 45 (goes even further, allowing short-notice "challenge inspections" on any facility of a State Party suspected of non-compliance).

218 See Mauricio Baker et al., *Verifying International Agreements on AI: Six Layers of Verification for Rules on Large-Scale AI Development and Deployment* (RAND Corp. Working Paper No. WR-A4077-1, 2025), https://www.rand.org/pubs/working_papers/WRA4077-1.html, . Also see Akash R. Wasil et al., *Verification methods for international AI agreements* (Aug. 28, 2024), https://arxiv.org/abs/2408.16074; Aaron Scher, David Abecassis, Peter Barnett, & Brian Abeyta, *An International Agreement to Prevent the Premature Creation of Artificial Superintelligence* (Nov. 13, 2025), https://arxiv.org/pdf/2511.10783, 31; Hamza Chaudry and James Petrie, *Innovations to Advance Compute Security - A Geo-Fencing Proposal* (forthcoming 2026).

219 See Hamza Chaudry and James Petrie, *Innovations to Advance Compute Security - A Geo-Fencing Proposal* (forthcoming 2026).

220 See Aaron Scher, David Abecassis, Peter Barnett, & Brian Abeyta, *An International Agreement to Prevent the Premature Creation of Artificial Superintelligence* (Nov. 13, 2025), https://arxiv.org/pdf/2511.10783, 32, 40-41; Christina Krawec, *Mapping IAEA Verification Tools to International AI Governance: A Mechanism-by-Mechanism Analysis* (Sept. 2025), https://simoninstitute.ch/blog/post/mapping-iaea-verification-tools-to-international-ai-governance-a-mechanism-by-mechanism-analysis.

221 See Matthijs Maas and José Jaime Villalobos, *International AI institutions: A literature review of models, examples, and proposals*, Inst. for L. & AI (Sept. 2023), https://law-ai.org/international-ai-institutions/, 25-26.

222 See Frank Ryan and Tereza Zoumpalova, *Why Whistleblowers Are Critical for AI Governance*, The Future Society (July 24, 2025), https://thefuturesociety.org/ai-whistleblowers.

223 S.B. 53, 2025–2026 Leg., Reg. Sess. (Cal. 2025), sec 1107.1.(a).

224 See OpenAI, *OpenAI Raising Concerns Policy* (Oct. 2024), https://cdn.openai.com/policies/raising-concerns-policy-blog-copy-202410.pdf.

225 Ben Harack et al., *Verification for International AI Governance* (AI Governance Initiative July 2025), https://aigi.ox.ac.uk/wp-content/uploads/2025/07/Verification_for_International_AI_Governance.pdf. All errors in the representation of this proposal for the purpose of these Articles are exclusively attributable to the lead author of this document. It is recommended that readers refer to the original paper for a more accurate and detailed description and justification of the proposal.

226 Confidential computing features are already available on some of the leading AI chips, so there is potential for a rapid rollout of this approach to govern an important portion of AI compute. Ben Harack et al., *Verification for International AI Governance* (AI Governance Initiative July 2025), https://aigi.ox.ac.uk/wp-content/uploads/2025/07/Verification_for_International_AI_Governance.pdf, 13.

227 Ben Harack et al., *Verification for International AI Governance* (AI Governance Initiative July 2025), https://aigi.ox.ac.uk/wp-content/uploads/2025/07/Verification_for_International_AI_Governance.pdf, 12.

228 Ben Harack et al., *Verification for International AI Governance* (AI Governance Initiative July 2025), https://aigi.ox.ac.uk/wp-content/uploads/2025/07/Verification_for_International_AI_Governance.pdf, 13.

229 Ben Harack et al., *Verification for International AI Governance* (AI Governance Initiative July 2025), https://aigi.ox.ac.uk/wp-content/uploads/2025/07/Verification_for_International_AI_Governance.pdf, 84. Combining both mutual local control with a relatively neutral host state is desirable for neutral data centers that perform sensitive verification operations. Given the extreme sensitivity of the operations that would be undertaken by a verification facility, choosing its location might be a key part of the negotiations for an international agreement.

230 Ben Harack et al., *Verification for International AI Governance* (AI Governance Initiative July 2025), https://aigi.ox.ac.uk/wp-content/uploads/2025/07/Verification_for_International_AI_Governance.pdf, 63-64. Also see Sella Nevo et al., *Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models*, RAND Corporation (30 May 2024), https://www.rand.org/pubs/research_reports/RRA2849-1.html.

231 Ben Harack et al., *Verification for International AI Governance* (AI Governance Initiative July 2025), https://aigi.ox.ac.uk/wp-content/uploads/2025/07/Verification_for_International_AI_Governance.pdf, 66.

232 Ben Harack et al., *Verification for International AI Governance* (AI Governance Initiative July 2025), https://aigi.ox.ac.uk/wp-content/uploads/2025/07/Verification_for_International_AI_Governance.pdf, 14.

233 For other proposals, see Onni Aarne, Tim Fist & Caleb Withers, *Secure, Governable Chips: Using On-Chip Mechanisms to Manage National Security Risks from AI & Advanced Computing* (Ctr. for a New Am. Sec. 2024), https://s3.us-east-1.amazonaws.com/files.cnas.org/documents/CNAS-Report-Tech-Secure-Chips-Jan-24-finalb.pdf; Gabriel Kulp and others, *Hardware-Enabled Governance Mechanisms: Developing Technical Solutions to Exempt Items Otherwise Classified Under Export Control Classification Numbers 3A090 and 4A090*, RAND Corporation (Jan. 2024), https://www.rand.org/pubs/working_papers/WRA3056-1.html; James Petrie et al., *Flexible Hardware-Enabled Guarantees for AI Compute*, arXiv:2506.15093 (June 18, 2025), https://arxiv.org/abs/2506.15093; Mauricio Baker et al., *Verifying International Agreements on AI: Six Layers of Verification for Rules on Large-Scale AI Development and Deployment* (RAND Corp. Working Paper No. WR-A4077-1, 2025), https://www.rand.org/pubs/working_papers/WRA4077-1.html; Aaron Scher & Lisa Thiergart, *Mechanisms to Verify International Agreements About AI Development* (Mach. Intel. Rsch. Inst., Nov. 27, 2024), https://techgov.intelligence.org/research/mechanisms-to-verify-international-agreements-about-ai-development; U.N. Sci. Advisory Bd., *Verification of Frontier AI* (June 2025), https://www.un.org/scientific-advisory-board/sites/default/files/2025-06/verification_of_frontier_ai.pdf.

234 James Petrie, *Near-Term Enforcement of AI Chip Export Controls Using A Firmware-Based Design for Offline Licensing*, arXiv:2404.18308 (Apr. 28, 2024); Onni Aarne, Tim Fist & Caleb Withers, *Secure, Governable Chips: Using On-Chip Mechanisms to Manage National Security Risks from AI & Advanced Computing* (Ctr. for a New Am. Sec. 2024), https://s3.us-east-1.amazonaws.com/files.cnas.org/documents/CNAS-Report-Tech-Secure-Chips-Jan-24-finalb.pdf; Gabriel Kulp and others, *Hardware-Enabled Governance Mechanisms: Developing Technical Solutions to Exempt Items Otherwise Classified Under Export Control Classification Numbers 3A090 and 4A090*, RAND Corporation (Jan. 2024), https://www.rand.org/pubs/working_papers/WRA3056-1.html; James Petrie et al., *Flexible Hardware-Enabled Guarantees for AI Compute*, arXiv:2506.15093 (June 18, 2025), https://arxiv.org/abs/2506.15093. All errors in the representation of these proposals for the purpose of these Articles are exclusively attributable to the lead author of this document. It is recommended that readers refer to the original papers for a more accurate and detailed description and justification of the proposal.

235 James Petrie, *Near-Term Enforcement of AI Chip Export Controls Using A Firmware-Based Design for Offline Licensing*, arXiv:2404.18308 (Apr. 28, 2024), 2.

236 James Petrie, *Near-Term Enforcement of AI Chip Export Controls Using A Firmware-Based Design for Offline Licensing*, arXiv:2404.18308 (Apr. 28, 2024), 2.

237 James Petrie et al., *Flexible Hardware-Enabled Guarantees for AI Compute*, arXiv:2506.15093 (June 18, 2025), https://arxiv.org/abs/2506.15093 https://arxiv.org/abs/2506.03409, 7-8, 35.

Draft Articles on the Safe Development and Diffusion of Artificial Intelligence

global-governance.ai