

TREATY ON ARTIFICIAL INTELLIGENCE SAFETY AND COOPERATION

Recognizing the existential risks posed by advanced artificial intelligence systems to humanity,

Aware of the immense benefits that artificial intelligence may bring to humanity and the shared desire to foster its responsible development,

Recognizing that the safe and advanced development of AI technologies should be pursued collaboratively, maximizing benefits to all and reducing inter-State AI race dynamics,

Acknowledging that AI safety constitutes a global public good and requires robust international cooperation and shared regulatory efforts,

Committing to a cooperative framework that ensures the safe, responsible, and sustainable development of AI technologies in accordance with the highest safety and ethical standards,

Cognizant that AI technologies can only yield the greatest benefits to humanity if developed and governed in accordance with comprehensive safety, ethical, and regulatory guidelines,

The signatories of this Treaty hereby agree to the following:

Article 1: Definitions

“API” means application programming interface.

“Artificial intelligence (AI)” means software and/or hardware that can learn to solve complex problems, make predictions or undertake tasks that require human-like sensing (such as vision, speech, and touch), perception, cognition, planning, learning, communication, or physical action.

“IASC” means International AI Safety and Cooperation Commission.

“JAISL” means Joint AI Safety Laboratory.

“FLOP” means floating-point operation.

“Large Training Run” means a training run of an AI model requiring a cumulative total of more than 10^{23} FLOP.

“Large Computing Clusters” means any aggregation of computational resources, including distributed or otherwise interconnected systems, capable of conducting a training run with a performance of 3×10^{17} FLOP per second.

Article 2: Objectives

The objectives of this Treaty are to

1. Prevent the development of AI systems that pose existential risks to humanity.
2. Promote transparency and cooperation between States Parties in the field of AI research and development.
3. Encourage the ethical and responsible use of AI technologies for the betterment of humanity.
4. Facilitate the development and availability of safe AI models.

Article 3: Ban on Large Training Runs

1. States Parties shall prohibit Large Training Runs under their jurisdiction or control and shall not assist, encourage or induce, in any way, anyone to engage in conducting Large Training Runs, with the exception of training runs conducted by the Joint AI Safety Laboratory.

Article 4: Closure of Large Computing Clusters

1. The Parties shall dismantle or repurpose Large Computing Clusters under their jurisdiction or control and shall not assist, encourage or induce, in any way, anyone to build, acquire or have control over Large Computing Clusters for AI development.
2. The Parties shall encourage the use of existing computational resources for projects that contribute to social, economic, and environmental benefits.

Article 5: National implementation

1. Each State Party shall adopt the necessary measures to implement its obligations under this Treaty.
2. Each State Party shall take all appropriate legal, administrative and other measures, including the imposition of penal sanctions, to prevent and suppress any activity prohibited under this Treaty undertaken by persons or on territory under its jurisdiction or control.
3. States Parties shall take special measures to, individually and in collaboration, identify, monitor, deter, and prevent potentially hazardous activities by State or non-State actors prohibited under this Treaty. These measures may include, but are not limited to:
 - (a) Developing a framework for responsible oversight, information collection, and intelligence sharing focused on AI safety which includes the use of established communication channels, joint threat assessment mechanisms, and collaboration with other States and international organizations, while at all times complying with domestic and international rules on privacy and data protection.
 - (b) Diplomatic pressure and law enforcement cooperation.

Article 6: International AI Safety and Cooperation Commission

1. The International AI Safety and Cooperation Commission (IASC) is hereby established.
2. All States Parties to this Convention shall be members of the Organization.
3. States Parties shall submit a monthly report to the IASC on their efforts to implement this Treaty. The report shall include the following information:
 - (a) A list of facilities and organizations working on the development of AI;
 - (b) A list of facilities and organizations working on AI safety;
 - (c) Information on training runs surpassing a cumulative total of 10% of the computation for a Large Training Run being conducted under the State Party's jurisdiction or control, including:
 - (i) The amount of computation (FLOP) and graphical processing units, or their equivalent in other hardware, being used in each run;

- (ii) A general statement on the purpose of each training run and the expected outcome; and
- (iii) A detailed account of the safety measures that are being applied in each training run.
- (d) A list of large computing clusters under their jurisdiction or control and progress on dismantling them;
- (e) Updates on AI safety research conducted by State or State-sponsored programmes under their jurisdiction or control; and
- (f) Incidents of non-compliance with this Treaty or activities suspected to potentially contravene this Treaty, within or outside a State Party's jurisdiction or control.

4. The functions of the IASC are to:

- (a) Monitor compliance with this Treaty by conducting inspections and audits of facilities and organizations under States Parties' jurisdiction or control working on the development of AI, in accordance with verification guidelines the IASC shall elaborate no later than 90 days after the entry into force of this Treaty;
- (b) Promote AI safety research;
- (c) At its discretion and in consultation with the JAISL, lower the limits for Large Training Runs or Large Computing Clusters established in Article 1 based on advancements that allow for more efficient use of compute resources in AI development;
- (d) At its discretion and in consultation with the JAISL, lower the limits for training runs conducted by the JAISL under Article 7.3(b);
- (e) At its discretion and in consultation with the JAISL, grant limited exceptions to the Large Training Run limit for specific narrow model types, provided that such exceptions adhere to the highest standards of safety and ethical use. These exceptions shall be subject to regular review and reassessment to ensure continued compliance with safety standards;
- (f) Make recommendations to States Parties to prohibit certain training runs that, despite requiring a total amount of computation below the limit for Large Training Runs, the IASC judges may nevertheless present an existential risk to humanity; and
- (g) Act as this Treaty's Secretariat and Depositary, accepting instruments of ratification or withdrawal, making arrangements for conferences of the Parties, compiling and communicating reports submitted to it, and providing States Parties with services as required.

5. Each State Party shall, upon entry into force of this Treaty, provide access to facilities and organizations under its jurisdiction or control working on the development of AI to the IASC for the purpose of monitoring compliance with this Treaty.

6. The IASC 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.

Article 7: Joint AI Safety Laboratory

1. The Joint AI Safety Laboratory (JAISL) is hereby established.

2. The JAISL is not subject to the ban on large training runs established in Article 3 but may not surpass the limits for large training runs indicated in this Article or established by the IASC.

3. The functions of the JAISL are to:

- (a) Conduct research, development, and innovation in the field of AI safety, with the primary goal of minimizing existential risks associated with AI technologies and prioritizing a solution to the alignment problem;
- (b) Regularly developing new AI models that each utilize a cumulative total of no more than 2.5×10^{25} FLOP for training while ensuring those models adhere to the highest safety and ethical standards, are beneficial to society, and meet the individual priorities of States Parties;
- (c) Provide access to the APIs of these models, with States Parties having the authority to set the pricing of the APIs for their respective countries, potentially raising revenues for purposes such as mitigating the disruptive economic effects of AI;
- (d) Collaborate with national and international research institutions, academia, and the private sector to foster innovation, advance the understanding and implementation of AI safety measures, and ensure that the JAISL remains at the forefront of AI safety and technology; and
- (e) Facilitate capacity building and training programs for researchers, policymakers, and other stakeholders to enhance global understanding of AI safety and promote responsible AI development.

4. The JAISL shall operate under the supervision of the IASC and shall have a robust institutional design consisting of an executive board and an advisory committee of experts. The decision-making processes of the JAISL shall ensure transparency and accountability.

5. The JAISL shall establish a system of checks and balances, which includes clearly defined roles and responsibilities for different stakeholders and their relationships with one another. This system shall help minimize mistakes and promote safety in the development and deployment of AI models.

6. The JAISL shall be committed to open and transparent research practices, sharing the results of its work, best practices, and breakthroughs in AI safety with the international community as long as the information it releases does not pose a hazard.

Article 8: Financial provisions

The funds required for the operation of this Treaty, including those for the functioning of the IASC and the JAISL, shall be charged exclusively against contributions from States Parties in accordance with a tiered contribution system based on each Party's economic capacity and access to resources. This system shall ensure that all Parties can participate in the Treaty's operation, while still providing sufficient funding, resources, and personnel to effectively support the activities of the IASC and the JAISL.

Article 9: Amendments

1. Any State Party may propose amendments to this Treaty.

2. The proposed amendments shall be considered and may be adopted only by a conference of the Parties organized by the IASC.

3. The text of any proposed amendment to this Treaty shall be communicated to States Parties by the IASC at least three months before the conference at which it is proposed for adoption.

4. States Parties shall make every effort to reach agreement on any proposed amendment to the Convention by consensus. If all efforts at consensus have been exhausted, and no agreement reached,

the amendment shall as a last resort be adopted by a three-fourths majority vote of States Parties present and voting at the meeting. The adopted amendment shall be circulated by the IASC to all States Parties for their acceptance.

5. Amendments shall enter into force for all States Parties 30 days after deposit of the instruments of ratification, acceptance, accession, or approval by all those States Parties which cast a positive vote at the amendment conference.

Article 10: Settlement of disputes

1. Any disputes arising from the interpretation or implementation of the Treaty shall be resolved through peaceful means, including negotiation, mediation, or arbitration.

2. The IASC may contribute to the settlement of the dispute, including by offering its good offices, calling upon the States Parties concerned to start the settlement procedure of their choice and recommending a time limit for any agreed procedure

Article 11: Signature, Entry into Force, Accession and Withdrawal

1. This Treaty shall be open to all States for signature before its entry into force.

2. This Treaty shall enter into force 90 days after the date of deposit of the instruments of ratification, acceptance, accession, or approval have been deposited by both the United States of America and the People's Republic of China.

3. Any State which does not sign this Treaty before its entry into force and This Treaty shall be open to accession by any State that shares the objectives and principles set forth in Article 2 of this Treaty may accede to it any time thereafter.

4. States Parties have the right to withdraw from this Treaty by giving notice of such withdrawal to the IASC. Such withdrawal will only take effect 6 months after the date of receipt of the notification of withdrawal by the IASC.

Article 12: Authentic Texts

This Treaty, of which the Arabic, Chinese, English, French, Russian, and Spanish texts are equally authentic, shall be deposited in the archives of the IASC. Duly certified copies of this Treaty shall be transmitted by the IASC to all States Parties.

DONE at [place], this [day] day of [month], [year].

IN WITNESS WHEREOF the undersigned, being duly authorized thereto by their respective Governments, have signed this Treaty.