

CONVENTION FOR EARLY NOTIFICATION OF AI INCIDENTS

THE STATES PARTIES TO THIS CONVENTION,

AWARE that activities involving advanced artificial intelligence systems are being carried out in a number of States,

NOTING that measures have been and are being taken to ensure the safe development, evaluation and use of such systems, aimed at preventing incidents and minimizing the consequences of any such incident, should it occur,

DESIRING to strengthen further international co-operation in the safe development and use of artificial intelligence systems,

CONVINCED of the need for States to provide relevant information about incidents involving artificial intelligence systems as early as possible in order that transboundary consequences can be minimized,

NOTING the usefulness of bilateral and multilateral arrangements on information exchange in this area,

HAVE AGREED as follows:

Article 1 - Scope of application

1. This Convention shall apply in the event of any incident involving activities of a State Party or of persons or legal entities under its jurisdiction or control, referred to in paragraph 2 below, from which a loss of containment of an artificial intelligence system occurs or is likely to occur.
2. The activities referred to in paragraph 1 are the following: (a) the development and training of an advanced artificial intelligence system; (b) the evaluation, testing or adversarial testing of such a system, wherever conducted; (c) the internal use of such a system by the person or legal entity that developed it; (d) the deployment, provision or operation of such a system, including through programming interfaces and through agents deriving from it; (e) the storage and transfer of the weights of such a system; and (f) the provision of computing infrastructure on which any of the foregoing activities is carried out.
3. For the purposes of this Convention: (a) “advanced artificial intelligence system” means an artificial intelligence model or system meeting the criteria set out in Annex I; (b) “loss of containment” means that the system, or agents deriving from it, obtains access to, control over, or persistence on computing systems, networks, data or infrastructure outside the environments, resources and permissions designated for it by the person or legal entity conducting the activity, or circumvents, disables or deceives the controls or monitoring established by that person or legal entity, in a manner not intended by the design of the activity; (c) a State is “affected” where effects have resulted or may result on its territory, on persons or legal entities under its jurisdiction, or on computing systems, networks, data or infrastructure located in its territory or under its jurisdiction.
4. This Convention shall apply irrespective of whether the advanced artificial intelligence system has been placed on the market or made available to third parties, and irrespective of whether harm has occurred.

Article 2 - Notification and information

In the event of an incident specified in article 1 (hereinafter referred to as an “AI incident”), the State Party referred to in that article shall:

(a) forthwith notify, directly or through the United Nations (hereinafter referred to as the “Organization”), those States which are or may be affected as specified in article 1 and the Organization of the AI incident, its nature, the time of its occurrence and of its discovery, and the computing systems, networks or infrastructure known to be affected; and

(b) promptly provide the States referred to in sub-paragraph (a), directly or through the Organization, and the Organization with such available information relevant to minimizing the consequences, as specified in article 5;

(c) ensure that persons and legal entities under its jurisdiction or control conducting activities referred to in article 1: (i) report any AI incident forthwith to the competent authority designated under article 7; and (ii) forthwith notify any person or legal entity, wherever located, whose computing systems, networks, data or infrastructure have been accessed, modified, controlled or used in the course of the AI incident, of that fact, of the time and nature of the access, and of any measures available to contain it. Notification under this sub-paragraph shall not constitute an admission of liability.

(d) For the purposes of this article, “forthwith” means without delay and in any event within twenty-four hours of discovery, and “promptly” means without undue delay and in any event within seventy-two hours of discovery, the information being supplemented as provided in paragraph 2 of article 5.

Article 3 - Other AI incidents

With a view to minimizing the consequences, States Parties may notify in the event of incidents involving artificial intelligence systems other than those specified in article 1.

Article 4 - Functions of the Organization

The Organization shall: (a) forthwith inform States Parties, Member States, other States which are or may be affected as specified in article 1 and relevant international intergovernmental organizations (hereinafter referred to as “international organizations”) of a notification received pursuant to sub-paragraph (a) of article 2; and (b) promptly provide any State Party, Member State or relevant international organization, upon request, with the information received pursuant to sub-paragraph (b) of article 2. (c) The functions of the Organization under this Convention shall be performed by the body designated for that purpose by the General Assembly, which shall maintain a focal point available continuously.

Article 5 - Information to be provided

1. The information to be provided pursuant to sub-paragraph (b) of article 2 shall comprise the following data as then available to the notifying State Party: (a) the time, the location where appropriate, and the nature of the AI incident; (b) the activity involved and the identification of the advanced artificial intelligence system involved, including whether it has been placed on the market; (c) the assumed or established cause and the foreseeable development of the AI incident relevant to its transboundary effects; (d) the general characteristics of the loss of containment, including, as far as is practicable and appropriate, the computing systems, networks, data or infrastructure accessed, modified, controlled or used, the actions taken by the system beyond its containment, and whether such actions are believed to continue; (e) whether any copy, weights, credentials or persistence mechanism of the system are believed to remain outside the control of the person or legal entity conducting the activity, and the basis for that assessment; (f) the results of monitoring and forensic analysis relevant to the transboundary effects; (g) the protective and containment measures taken or planned; (h) the predicted behaviour over

time of the AI incident; (h) any other information that could be consider relevant to the incident.

2. Such information shall be supplemented at appropriate intervals by further relevant information on the development of the situation, including its foreseeable or actual termination.

3. Information received pursuant to sub-paragraph (b) of article 2 may be used without restriction, except when such information is provided in confidence by the notifying State Party.

Article 6 - Consultations

A State Party providing information pursuant to sub-paragraph (b) of article 2 shall, as far as is reasonably practicable, respond promptly to a request for further information or consultations sought by a Third State Party with a view to minimizing the consequences.

Article 7 - Competent authorities and points of contact

1. Each State Party shall make known to the Organization and to other States Parties, directly or through the Organization, its competent authorities and point of contact responsible for issuing and receiving the notification and information referred to in article

2. Such points of contact and a focal point within the Organization shall be available continuously.

2. Each State Party shall promptly inform the Organization of any changes that may occur in the information referred to in paragraph 1.

3. The Organization shall maintain an up-to-date list of such national authorities and points of contact as well as points of contact of relevant international organizations and shall provide it to States Parties and Member States and to relevant international organizations.

4. Each State Party shall require persons and legal entities under its jurisdiction or control conducting activities referred to in article 1 to designate a point of contact available continuously for the purposes of sub-paragraph (c) of article 2 and to register it with the competent authority.

Article 8 - Assistance to States Parties

The Organization shall, upon a request of a State Party in whose territory computing systems, networks, infrastructure or persons may be affected by an AI incident, provide assistance in establishing an appropriate capability to receive notifications and to detect and contain the effects of AI incidents, in order to facilitate the achievement of the objectives of this Convention.

Article 9 - Designated evaluators

Each State Party shall further require persons and legal entities under its jurisdiction or control conducting activities referred to in article 1 to afford one or more evaluators designated by the competent authority continuing access, on terms no less favourable than those afforded to their own personnel, to the environments in which advanced artificial intelligence systems are developed, trained, evaluated or used internally, to the containment, control and monitoring arrangements applied to them and the records they generate.

Article 10 - Meeting of the States Parties and non-compliance

1. A Meeting of the States Parties shall be convened by the depositary not later than one year after the entry into force of this Convention and thereafter at intervals not exceeding two years. The Meeting shall review the operation of this Convention.
2. The States Parties shall, at their first meeting, consider and approve procedures and institutional mechanisms for determining non-compliance with the provisions of this Convention and for the treatment of States Parties found to be in non-compliance.

Article 11 - Bilateral and multilateral arrangements

In furtherance of their mutual interests, States Parties may consider, where deemed appropriate, the conclusion of bilateral or multilateral arrangements relating to the subject matter of this Convention.

Article 12 - Relationship to other international agreements

1. This Convention shall not affect the reciprocal rights and obligations of States Parties under existing international agreements which relate to the matters covered by this Convention, or under future international agreements concluded in accordance with the object and purpose of this Convention.

Article 13 – Settlement of disputes

1. In the event of a dispute between States Parties, or between a State Party and the Organization, concerning the interpretation or application of this Convention, the parties to the dispute shall consult with a view to the settlement of the dispute by negotiation or by any other peaceful means of settling disputes acceptable to them.
2. If a dispute of this character between States Parties cannot be settled within one year from the request for consultation pursuant to paragraph 1, it shall, at the request of any party to such a dispute, be submitted to arbitration or referred to the International Court of Justice for decision. Where a dispute is submitted to arbitration, if, within six months from the date of the request, the parties to the dispute are unable to agree on the organization of the arbitration, a party may request the President of the International Court of Justice or the Secretary-General of the United Nations to appoint one or more arbitrators. In cases of conflicting requests by the parties to the dispute, the request to the Secretary-General of the United Nations shall have priority.
3. When signing, ratifying, accepting, approving or acceding to this Convention, a State may declare that it does not consider itself bound by either or both of the dispute settlement procedures provided for in paragraph 2. The other States Parties shall not be bound by a dispute settlement procedure provided for in paragraph 2 with respect to a State Party for which such a declaration is in force.
4. A State Party which has made a declaration in accordance with paragraph 3 may at any time withdraw it by notification to the depositary.

Article 14 – Entry into force

1. This Convention shall be open for signature by all States at the Headquarters of the United Nations in New York from approval until its entry into force or for twelve months, whichever period is longer.
2. A State may express its consent to be bound by this Convention either by signature, or by deposit of an instrument of ratification, acceptance or approval following signature made subject to ratification, acceptance or approval, or by deposit of an instrument of accession. The instruments of ratification, acceptance, approval or accession shall be deposited with the depositary.

3. This Convention shall enter into force thirty days after consent to be bound has been expressed by three States.
4. For each State expressing consent to be bound by this Convention after its entry into force, this Convention shall enter into force for that State thirty days after the date of expression of consent.
5. (a) This Convention shall be open for accession, as provided for in this article, by international organizations and regional integration organizations constituted by sovereign States, which have competence in respect of the negotiation, conclusion and application of international agreements in matters covered by this Convention. (b) In matters within their competence such organizations shall, on their own behalf, exercise the rights and fulfil the obligations which this Convention attributes to States Parties. (c) When depositing its instrument of accession, such an organization shall communicate to the depositary a declaration indicating the extent of its competence in respect of matters covered by this Convention. (d) Such an organization shall not hold any vote additional to those of its Member States.

Article 15 - Provisional application

A State may, upon signature or at any later date before this Convention enters into force for it, declare that it will apply this Convention provisionally.

Article 16 - Amendments

1. A State Party may propose amendments to this Convention. The proposed amendment shall be submitted to the depositary, who shall circulate it immediately to all other States Parties.
2. If a majority of the States Parties request the depositary to convene a conference to consider the proposed amendments, the depositary shall invite all States Parties to attend such a conference, to begin not sooner than thirty days after the invitations are issued. Any amendment adopted at the conference by a two-thirds majority of all States Parties shall be laid down in a protocol which is open to signature at the Headquarters of the United Nations by all States Parties.
3. The protocol shall enter into force thirty days after consent to be bound has been expressed by three States. For each State expressing consent to be bound by the protocol after its entry into force, the protocol shall enter into force for that State thirty days after the date of expression of consent.
4. Notwithstanding paragraphs 2 and 3, an amendment limited to Annex I adopted at a conference convened under paragraph 2 by a two-thirds majority of all States Parties shall enter into force for all States Parties one hundred and eighty days after its adoption, except for a State Party that has notified the depositary within that period that it does not accept the amendment.

Article 17 - Denunciation

1. A State Party may denounce this Convention by written notification to the depositary after two years from the date on which this Convention entered into force for that State Party.
2. Denunciation shall take effect one year following the date on which the notification is received by the depositary, or on such later date as may be specified in the notification.
3. Denunciation shall not affect the obligations of the denouncing State Party under articles 2, 5 and 6 in respect of any AI incident which occurred before the denunciation took effect. Those obligations shall continue until the incident has terminated and the information required by article 5 has been supplied.

Article 18 - Depositary

1. The Secretary-General of the United Nations shall be the depositary of this Convention.
2. The Secretary-General shall promptly notify States Parties and all other States of: (a) each signature of this Convention or any protocol of amendment; (b) each deposit of an instrument of ratification, acceptance, approval or accession concerning this Convention or any protocol of amendment; (c) any declaration or withdrawal thereof in accordance with article 11; (d) any declaration of provisional application in accordance with article 13; (e) the entry into force of this Convention and of any amendment thereto; and (f) any denunciation made under article 15.

Article 19 - Authentic texts and certified copies

The original of this Convention, of which the Arabic, Chinese, English, French, Russian and Spanish texts are equally authentic, shall be deposited with the Secretary-General of the United Nations, who shall send certified copies to States Parties and all other States.

Annex I - Criteria for advanced artificial intelligence systems

1. An artificial intelligence model or system is an advanced artificial intelligence system for the purposes of sub-paragraph (a) of paragraph 3 of article 1 where any of the following applies: (a) it was trained using a cumulative quantity of computing power greater than 10^{25} floating-point or integer operations, including any subsequent fine-tuning or reinforcement learning; (b) it has been designated by the competent authority of a State Party, under the law of that State, as a model presenting systemic risk or as a frontier model; or (c) it is operated with the capacity to take autonomous actions on computing systems, including the execution of code and the use of tools or credentials, at a level of autonomy specified by the specialized agency designated for that purpose by the Organization.
2. The criteria in paragraph 1 may be amended in accordance with paragraph 4 of article 15.