# Cover Letter

This treaty is absolutely necessary to secure the future of humanity. The rapid progression of artificial intelligence research and development poses among other risks a significant existential threat to the human race.

This treaty is written by Daniel Marns and Ben Anderson, undergraduates at the Australian National University. The approach we took was to draw from other international treaties - such the Nuclear Non-Proliferation Treaty - and perform a literature review of academic work relevant to artificial intelligence and treaties. A treaty is more likely to be successful if it does not break massively from the accepted status quo. The central purpose of the treaty is to call for a moratorium of large-scale AI capabilities research and development, as such the more extreme recommendations outlined in Eliezer Yudkowsky’s TIME article - such as ‘be willing to destroy a rogue datacenter by airstrike’ - are not forgotten but purposely excluded to increase the chance of achieving the central purpose.

When writing this treaty it is important to consider the context this treaty could possibly be used in. We chose to structure and write the treaty using the structure of other United Nations international treaties. The reason is as this would be the most likely context that the treaty would be introduced on the international stage, for example it could be taken directly to representatives of a smaller country - perhaps local in the Australasian region - who could then take it directly to the international sphere through the processes of the United Nations. By using ideas, structure and language from existing treaties about nuclear weapons and autonomous weapons we believe the treaty is more likely to be successful, every new concept introduced is going to be a fight, so keeping as much the same and focusing on whats most relevant is absolutely necessary.

We view it as important that the treaty expresses some focus on the day to day issues of AI, for example not only a hard cap on AI weapons and general AI super intelligence.

Our treaty includes an explanatory memorandum. These are **not intended** to be read as part of the treaty itself. Rather, they include discursive discussion on the purpose of particular provisions, and sources to further information. Indeed, such ‘guiding’ discussion would be available in a real treaty in the form of its *travaux preparatoires,* which can guide interpretation in accordance with article 32 of the Vienna Convention of the Law of Treaties.

## Treaty on the Moratorium of Large-Scale AI Capabilities Research and Development

*The States Parties to this Treaty,*

Determined to contribute to the realisation of the purposes and principles of the Charter of the United Nations,

Deeply concerned about the catastrophic humanitarian consequences that would result from any use of artificial general intelligence (AGI), and recognising the consequent need to completely eliminate further relevant research and design, which remains the only way to guarantee that AGI is not developed until it is definitively safe for it to do so.

Mindful of the risks posed by the potential development of AGI, including from any alignment problems, and emphasising that these risks concern the security of all humanity, and that all States share the responsibility to prevent any further development of large-scale AI capabilities,

Cognizant that the catastrophic consequences of AGI cannot be adequately addressed, transcend national borders, pose grave implications for human survival, the environment, socioeconomic development, the global economy and the health of current and future generations

Acknowledging the ethical imperatives for preventing large-scale AI capability research and development and the urgency of maintaining an AGI free world, which is a global public good of the highest order, serving both national and collective security interests,

Reaffirming the need for all States at all times to comply with applicable international law, including international humanitarian law and international human rights law,

Basing themselves on the principles and rules of international humanitarian law,

Reaffirming that any development of AGI would also be abhorrent to the principles of humanity and the dictates of public conscience,

Recognising that a legally binding moratorium of large-scale AI capabilities research and development constitutes an important contribution towards the maintenance of a world free of AGI, and determined to act towards that end,

Reaffirming that there exists an obligation to pursue in good faith and bring to a conclusion negotiations leading to an AGI moratorium in all its aspects under strict and effective international control,

Recognising that the equal, full and effective participation of both women and men is an essential factor for the promotion and attainment of sustainable peace and security, and committed to supporting and strengthening the effective participation of women in further AI safety research,

Recognising also the importance of peace and AI alignment education in all its aspects and of raising awareness of the risks and consequences of AGI for current and future generations, and committed to the dissemination of the principles and norms of this Treaty,

Stressing the role of public conscience in the furthering of the principles of humanity as evidenced by the call for the moratorium of large-scale AI capabilities research and design, and recognizing the efforts to that end undertaken by the United Nations, other international and regional organisations, non-governmental organisations, parliamentarians, field leaders and academics,

*Have agreed as follows:*

### Article I - Definitions

For the purposes of the present Convention:

1. The term "large-scale artificial intelligence" refers to any algorithmic system or a combination of such systems that uses computational methods derived from statistics or other mathematical techniques to carry out functions that are commonly associated with, or would otherwise require, human intelligence and that either assists or replaces the judgement of human decision-makers in carrying out those functions. Such functions include, but are not limited to, prediction, planning, classification, pattern recognition, organisation, perception, speech/sound/image recognition, text/sound/image generation, language translation, communication, learning, representation, and problem solving;
2. The term “artificial intelligence capabilities research and development” refers to the activities focused on advancing the capabilities of artificial intelligence systems;
3. The term “artificial intelligence specialised computer chips” refers to hardware components designed specifically to handle the computational demands of artificial intelligence systems workload;
4. The term ‘artificial general intelligence’ refers to highly autonomous, agentic variations of the systems described in Article I(1), capable of operating independent of any human input, with the ability to exceed human abilities in a large variety of tasks.

### Article II - Establishment of the Agency

1. There shall be a body entitled the International Artificial Intelligence Agency, herein referred to as “‘the Agency”.

### Article III - Objective

1. The Agency shall seek to ensure that the development of artificial intelligence is conducted in concordance with the relevant Articles in this statute.

### Article IV - Facilitative Functions

1. The Agency is authorised to;
   1. To encourage the exchange and training of artificial intelligence safety experts by State Parties.
   2. To foster the exchange of scientific and technical information on the safe use of artificial intelligence between State Parties.
   3. Provide the necessary equipment to State Party’s to facilitate the development of safe artificial intelligence systems.

### Article V - Other Functions

1. The Agency is authorised;
   1. To set the relevant computational parameter parameter threshold under which State Party’s are prohibited from testing artificial intelligence systems under Article XV(2).
   2. This threshold is, by default, to be set every 6 months, beginning three months after the initial entry into force. This can be reduced at the discretion of the Agency.
2. To ascertain what computational resources, including artificial intelligence specialised computer chips, are required to create advanced artificial intelligence systems.
3. To establish and administer safeguard agreements with State Parties designed to ensure that domestic facilities are not utilised to create artificial intelligence specialised computer chips.
   1. This may include, but is not limited to;
      1. Scheduled and unscheduled inspections enacted to ensure any such facility is conducted in a manner conducive to the effective applications of the safeguards provided for by the Agency.
      2. Remote monitoring programs enacted to ensure any such facility is conducted in a manner conducive to the effective applications of the safeguards provided for by the Agency.

### Article VI - Confidentiality

1. In any safeguards enacted under Article V, the Agency shall;
   1. Require only the minimum amount of data and information needed to ensure compliance with the relevant obligations under this Treaty.
   2. Take all requisite measures to ensure none of this information or data is protected from release or distribution.

### Article VII - State Obligations

1. Each State Party agrees to undertake relevant safeguards and monitoring provisions in accordance with an agreement between the state and the Agency made under Article V.

### Article VIII - Governance Structure

1. A General Conference consisting of representatives of all State Parties’ shall meet in regular annual session. In this Conference, State Parties shall vote on the membership of the Board of the Agency, herein referred to as the Board.
2. The Board shall be constituted as follows;
   1. There shall be 10 board seats.
   2. The United States of America and the People’s Republic of China shall hold positions on the Board of the Agency, if they are signatories to the Treaty.
   3. The remaining board seats shall be filled with State Parties on a vote of the General Conference.
3. The Board shall have authority to carry out the functions of the Agency in accordance with this Treaty.
4. The Board shall take all due measures to ensure the transition to the new Board after a General Conference does not impede on the ability of the Agency to conduct its operations.

### Article IX - Non-Compliance

1. In the event of non-compliance and failure by the non-compliant State or States to take requested corrective steps within a reasonable time, the Agency suspend or terminate assistance and withdraw any materials and equipment made available by the Agency or a member in furtherance of the project.
2. At the vote of all members of the Board, the Agency may remove the non-compliant state from the Board.

### Article X - Termination

1. The State Parties may, at a meeting of the kind set forth in Article VII(1), agree to terminate the treaty through a vote.
   1. All State Parties must vote in favour of terminating the treaty.
   2. It is intended this treaty shall only be terminated whereby all State Parties agree that as of widespread, rigorously tested technological safety advancements, it is safe to resume large scale AI capabilities research and development.

### Article XI - Obligation to shut down research

1. Each State Party to the Treaty undertakes to accept safeguards, as set forth in an agreement to be negotiated and concluded with the Agency in accordance with the Statute of the Agency and the Agency's safeguards system, for the exclusive purpose of verification of the fulfilment of its obligations assumed under this Treaty with a view to preventing further large-scale AI capabilities research and design. Procedures for the safeguards required by this Article shall be followed with respect to the decommissioning of large GPU and TPU clusters where large-scale artificial intelligence systems are refined.
2. Each State Party to the Treaty undertakes not to:
   1. source or provide special computational material designed for use in specialised AI GPUs or TPUs, or;
   2. equipment or material especially designed or prepared for the development, use or production of large scale artificial intelligence systems.
3. States Party to the Treaty shall conclude agreements with the Agency to meet the requirements of this Article either individually or together with other States in accordance with the Statute of the Agency. Negotiation of such agreements shall commence within 180 days from the original entry into force of this Treaty. For States depositing their instruments of ratification or accession after the 180-day period, negotiation of such agreements shall commence not later than the date of such deposit. Such agreements shall enter into force not later than eighteen months after the date of initiation of negotiations.

### Article XII - Obligation to introduce domestic laws

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 to a State Party under this Treaty undertaken by persons or on territory under its jurisdiction or control.
3. Each State Party must ensure the measures taken criminalise on a national scale the development of any form of Artificial General Intelligence (AGI).

Article XIII - International cooperation and assistance

1. Each State Party shall cooperate with other States Parties to facilitate the implementation of this Treaty.
2. In fulfilling its obligations under this Treaty, each State Party shall have the right to seek and receive assistance, where feasible, from other States Parties.
3. Assistance under this Article may be provided, inter alia, through the United Nations system, international, regional or national organisations or institutions, non-governmental organisations or institutions, or on a bilateral basis.

### Article XIV - Regional treaties

1. Nothing in this Treaty affects the right of any group of States to conclude regional treaties in

in order to assure a moratorium on large-scale AI capabilities research and development.

### Article XV - Prohibition of further artificial intelligence development

1. Each State Party to the Treaty undertakes not to participate in any form of development, testing, or production of quantum computing technologies for the purpose of artificial intelligence research, design, development or application. This prohibition extends to both direct involvement in such activities and indirect involvement through provision of financial, technical, or material support.
2. Each State Party to the Treaty agrees to a general moratorium on large-scale AI capabilities research and development. This includes, but is not limited to, the training, testing, or application of artificial intelligence systems with computational capabilities exceeding a threshold as established by the Agency. This threshold is initially set to systems involving more than five-hundred million parameters, but may be revised in accordance with scientific and technical progress as determined by the Agency.
3. The Agency is charged with the responsibility of determining the appropriate measures for the verification of compliance with the obligations outlined in this Article, and for the enforcement of these measures in accordance with the Statute of the Agency.

### Article XVI - Meeting of States Parties

1. The States Parties shall meet regularly in order to consider and, where necessary, take decisions in respect of any matter with regard to the application or implementation of this Treaty, in accordance with its relevant provisions, and on further measures for a moratorium of large-scale AI capabilities research and development, including:
   1. The implementation and status of this Treaty;
   2. Measures for moratorium of large-scale artificial intelligence capabilities research and design, including additional protocols to this Treaty;
   3. Any other matters pursuant to and consistent with the provisions of this Treaty.
2. The first meeting of States Parties shall be convened by the Secretary-General of the United Nations within one year of the entry into force of this Treaty. Further meetings of States Parties shall be convened by the Secretary-General of the United Nations on a biennial basis, unless otherwise agreed by the States Parties. The meeting of States Parties shall adopt its rules of procedure at its first session. Pending their adoption, the rules of procedure of the United Nations conference to negotiate a legally binding instrument for a moratorium of large-scale artificial intelligence capabilities research and development shall apply.
3. Extraordinary meetings of States Parties shall be convened, as may be deemed necessary, by the Secretary-General of the United Nations, at the written request of any State Party provided that this request is supported by at least one third of the States Parties.
4. After a period of five years following the entry into force of this Treaty, the Secretary-General of the United Nations shall convene a conference to review the operation of the Treaty and the progress in achieving the purposes of the Treaty. The Secretary-General of the United Nations shall convene further review conferences at intervals of six years with the same objective, unless otherwise agreed by the States Parties.
5. States not party to this Treaty, as well as the relevant entities of the United Nations system, other relevant international organisations or institutions, regional organisations, shall be invited to attend the meetings of States Parties and the review conferences as observers.

### Article XVII - Costs

1. The costs of the meetings of States Parties, the review conferences and the extraordinary meetings of States Parties shall be borne by the States Parties and States not party to this Treaty participating therein as observers, in accordance with the United Nations scale of assessment adjusted appropriately.
2. The costs incurred by the Secretary-General of the United Nations in the circulation of declarations under Article 2, reports under Article 4 and proposed amendments under Article 10 of this Treaty shall be borne by the States Parties in accordance with the United Nations scale of assessment adjusted appropriately.

### Article XVIII - Amendments

1. At any time after the entry into force of this Treaty, any State Party may propose amendments to the Treaty. The text of a proposed amendment shall be communicated to the Secretary-General of the United Nations, who shall circulate it to all States Parties and shall seek their views on whether to consider the proposal. If a majority of the States Parties notify the Secretary-General of the United Nations no later than 90 days after its circulation that they support further consideration of the proposal, the proposal shall be considered at the next meeting of States Parties or review conference, whichever comes first.
2. A meeting of States Parties or a review conference may agree upon amendments which shall be adopted by a positive vote of a majority of two thirds of the States Parties. The Depositary shall communicate any adopted amendment to all States Parties.
3. The amendment shall enter into force for each State Party that deposits its instrument of ratification or acceptance of the amendment 90 days following the deposit of such instruments of ratification or acceptance by a majority of the States Parties at the time of adoption. Thereafter, it shall enter into force for any other State Party 90 days following the deposit of its instrument of ratification or acceptance of the amendment.

### Article XIX - Settlement of disputes

1. When a dispute arises between two or more States Parties relating to the interpretation or application of this Treaty, the parties concerned shall consult together with a view to the settlement of the dispute by negotiation or by other peaceful means of the parties’ choice in accordance with Article 33 of the Charter of the United Nations.
2. The meeting of States Parties 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, in accordance with the relevant provisions of this Treaty and the Charter of the United Nations.

### Article XX - Universality

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

### Article XXI - Signature

This Treaty shall be open for signature to all States at United Nations Headquarters in New York as from the 15th of July 2023.

### Article XXII - Ratification, acceptance, approval or accession

This Treaty shall be subject to ratification, acceptance or approval by signatory States. The Treaty shall be open for accession.

### Article XXIII - Entry into force

1. This Treaty shall enter into force 90 days after the twentieth instrument of ratification, acceptance, approval or accession has been deposited.
2. For any State that deposits its instrument of ratification, acceptance, approval or accession after the date of the deposit of the fiftieth instrument of ratification, acceptance, approval or accession, this Treaty shall enter into force 90 days after the date on which that State has deposited its instrument of ratification, acceptance, approval or accession.

### Article XXIV - Reservations

1. The Articles of this Treaty shall not be subject to reservations.

### Article XXV - Relationship with other agreements

The implementation of this Treaty shall not prejudice obligations undertaken by States Parties with regard to existing international agreements, to which they are parties, where those obligations are consistent with the Treaty.

### Article XXVI - Depositary

The Secretary-General of the United Nations is hereby designated as the Depositary of this Treaty.

### Article XXVII - Authentic texts

The Arabic, Chinese, English, French, Russian and Spanish texts of this Treaty shall be equally authentic.

# Explanatory Note

Article III; This was intended as an incentive mechanism for states to become and remain signatories.

Article IV(1)(b); A balance is required here between allowing for the threshold to be altered with improvements in data quality and algorithmic efficiency, whilst also not being too recurrent such as to be disruptive.

Article V(1)(b); A balance is required here between allowing for the threshold to be altered with improvements in data quality and algorithmic efficiency, whilst also not being too recurrent such as to be disruptive. Very recurrent resetting of the thresholds may result in uncertainty and thus disincentivize state parties from becoming signatories.

Article V(1)(b); This is to allow for the Agency to conduct investigations to determine what a reasonable initial threshold may be.

Article V(1)(b); This is included in case of dramatic improvements in algorithmic efficiency requiring substantial reduction in parameter thresholds to prevent AGI development.

Article V(2);As algorithmic efficiency grows, the types of computer chips and resources needed to create advanced artificial intelligence systems will grow. As such, the nature of the monitoring programs needed to limit AI training runs would change over time. There is also the potential for an overly broad definition to capture consumer-grade hardware. As such, any potential definition would need to be well scoped and specific. However, detailing the minutiae of such definitions would be far beyond the scope of a multilateral treaty. As such, art 3(1) seeks to vest this power in a flexible way to the agency itself.

Article V(3); Here, we focus on the governance of the physical chips used to run artificial intelligence systems. Monitoring chips is easier than monitoring the other primary vectors for improving AI, data and algorithms. This is because chips are a physical entity that can be tracked and monitored. The literature on the technical aspects of safeguards is undeveloped. As such, this particular task is left as a function for the agency, For more information on what this may look like, see <https://arxiv.org/pdf/2304.04123.pdf> and <https://arxiv.org/abs/2303.11341>

Article VIII(2)(b); These are widely considered to be the most important nations with regards to the progress of artificial intelligence. This provision is thus included for two reasons. Firstly, as an incentive for the USA and China to become signatories. Secondly, to ensure the expertise held by these nations is represented on the Board. This may disincentive smaller states from becoming signatories, however we have deemed the potential benefit from increasing the likelihood of the USA and China joining the treaty to outweigh this potential disadvantage.

Article X(1)(b): It is difficult to make this an actual *binding* condition on states terminating the treaty. Such a requirement may be invalid under the *Vienna Convention of the Law of Treaties.* As such, the phrasing here is non-determinative.