An International Appeal to Preserve the Possibility of Protection Against Catastrophic Artificial Intelligence Risks

AI Governance Watch (AIW)

26 November 2026

To the Secretary-General of the United Nations, political leaders, international organizations, trade unions, environmental organizations, scientific institutions, technology leaders, influential public figures and representatives of civil society

Artificial intelligence offers extraordinary opportunities for humanity. Its potential contributions to science, medicine, education, economic development and human welfare are immense. However, its rapidly increasing capabilities also raise fundamental questions about humanity's ability to protect itself against catastrophic consequences.

Recent developments suggest a growing recognition of these dangers, including among leading AI developers themselves.

Proposals for shared limits on technological advancement, independent safety evaluations and improved corporate governance are important contributions.

Nevertheless, these initiatives remain vulnerable to commercial competition, geopolitical rivalry and conflicting national security interests.

No individual company or country can ensure that other actors will exercise equivalent restraint.

The broader catastrophic risks

Catastrophic consequences could arise not only from the loss of control over AI systems, but also from their deliberate misuse, the destabilization of critical infrastructure and the economic, environmental and geopolitical consequences of unrestricted competition.

Even perfectly aligned AI could be dangerous if those controlling it instructed it to perform catastrophic actions.

Furthermore, the rapidly increasing electricity requirements of AI and its associated data centres threaten to increase energy costs, place additional pressure on electricity grids and exacerbate the climate crisis.

According to the International Energy Agency, electricity consumption by AI-focused data centres increased by 50% in 2025 and is projected to triple between 2025 and 2030.

The climate crisis is already a catastrophic threat to humanity. The international race to develop increasingly powerful AI must not be allowed to deepen that threat.

Competition over energy, computing infrastructure, semiconductors, capital and technological expertise could also intensify economic insecurity, industrial displacement, inequalities and international tensions.

Under certain circumstances, these developments could contribute to a major international war, potentially involving nuclear weapons.

Humanity could therefore face catastrophic consequences from the AI race even without AI itself becoming uncontrollable.

The need for a global constitutional architecture

The international community faces a challenge that cannot adequately be addressed through voluntary agreements or fragmented national regulation alone.

We propose that the United Nations initiate discussions concerning the establishment of a globally binding constitutional architecture governing potentially catastrophic AI capabilities.

Such an architecture would establish enforceable boundaries on the authority exercised by AI systems and by the companies, governments and other institutions controlling them.
It would also establish binding limitations on the authority of the international institutions responsible for enforcing these constraints.

The proposed framework would include:

- Internationally binding restrictions on the development, deployment and dissemination of potentially catastrophic AI capabilities.
- Independent verification without requiring complete disclosure of sensitive national or commercial information or mutual trust.
- Technical safeguards against the unilateral exercise of potentially catastrophic authority and the unilateral removal of protections.
- Reliable mechanisms for interrupting dangerous activities.
- International decision-making procedures without individual or collective great-power vetoes.
- Guaranteed representation of smaller countries through a quota system.
- Constitutional safeguards against the concentration or abuse of international enforcement authority.

AIW proposes a modified UN-based institutional structure involving a 14-member council, with decisions requiring a qualified majority of nine members.

The United States, China and Russia would have to accept that decisions could be adopted against their wishes, individually or collectively.

Smaller countries must have constitutionally guaranteed representation. Countries with limited influence over AI development may nevertheless suffer devastating consequences from its economic, environmental and geopolitical effects.

Within this global constitutional framework, specific technical, national and regional solutions could be developed incrementally.

The basis for international cooperation

The proposal does not require agreement on political systems, economic models or broader ideological questions.

Nor does it assume that governments or corporations will act primarily out of concern for humanity as a whole.
Its foundation is the recognition of a common existential self-interest: that no country, however powerful, can unilaterally preserve even a chance of any protection against the consequences of unconstrained catastrophic authority exercised elsewhere.

We recognize the extraordinary political and institutional difficulties involved.

Nevertheless, waiting until a catastrophe has demonstrated the inadequacy of existing arrangements may mean waiting until effective action is no longer possible.

An invitation to international action

We invite the recipients of this letter to:

1. Consider the attached proposal as a basis for international discussion.
2. Encourage the inclusion of binding constitutional constraints in forthcoming UN discussions on AI governance.
3. Support dialogue involving governments, scientific institutions, technology developers, trade unions, environmental organizations and civil society.
4. Explore the institutional, legal and technical feasibility of establishing enforceable international limits on potentially catastrophic AI capabilities.
5. Promote international consideration of the economic, environmental and geopolitical risks associated with unrestricted AI competition.

The fundamental principle must be:

No local actor may exercise unconstrained catastrophic authority, and no global institution established to enforce the constraints may itself possess unconstrained catastrophic authority.

The objective is not to guarantee humanity's safety. No institutional arrangement can provide such a guarantee.

It is to preserve a chance of any protection.

We believe that this objective deserves urgent and serious international consideration.

AI Governance Watch (AIW)

Dr Sarah Philipson

Associate Professor Emerita

ai-governance-watch.org

ai-governance-watch.com

***Attachment**: A Proposed Global Constitutional Architecture for Protection Against Catastrophic AI Risks*

***Reference**: International Energy Agency, Key Questions on Energy and AI, April 2026. Further references are provided in the attached institutional proposal.*