From Voluntary Safety Measures to a Global Constitutional Architecture

AI Governance Watch (AIW) | 26 November 2026

Recent developments in artificial intelligence have reinforced the need to move beyond fragmented national regulation, corporate self-governance and voluntary international cooperation.

They also suggest that the dangers associated with AI may be broader than is generally acknowledged.

The time has come to consider a fundamentally different approach to AI governance.

Recent developments

OpenAI's chief scientist, Jakub Pachocki, has proposed that competing AI developers agree on shared, measurable limits to technological advancement, allowing safety measures to keep pace with increasing AI capabilities. [1]

This is an important initiative. It recognizes that individual developers cannot easily exercise restraint when their competitors continue to advance.

Yet almost simultaneously, Reuters has reported that geopolitical competition with China is encouraging American political leaders to support the development and dissemination of open-weight AI models. According to Reuters, the US administration has excluded open-weight models from voluntary government safety testing. [2]

Open-weight models offer substantial benefits, including wider access, lower costs and reduced dependence on a small number of corporations. However, once models possessing potentially catastrophic capabilities are widely disseminated, subsequent restrictions may become practically impossible to enforce.

These developments illustrate a fundamental contradiction: while some leading developers recognize the need for shared limits, commercial and geopolitical pressures may undermine the possibility of establishing them.

Other recent developments reinforce this concern.

California's lawsuit against OpenAI raises questions about the adequacy of safeguards governing autonomous AI systems. The allegations remain to be adjudicated, but the case highlights the importance of independent oversight and external accountability. [3]

Anthropic's unusual corporate governance arrangements represent a serious attempt to incorporate humanity's long-term interests into corporate decision-making. Nevertheless, their effectiveness in resolving fundamental conflicts between commercial interests and safety remains unproven. [4]

Neither responsible corporate leadership nor technical safeguards alone can provide an adequate solution.

Catastrophic risks beyond the loss of control

The dangers associated with AI are not confined to the possibility that increasingly autonomous systems might escape human control.

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

Furthermore, recent disruptions to critical infrastructure have highlighted the fragility of interconnected technological systems. Although the reported outages cannot generally be attributed to AI, AI-assisted discovery of software vulnerabilities may increase pressures on already overstretched cybersecurity systems. [5]

Electricity consumption and the climate crisis

One of the most fundamental threats associated with the rapid expansion of AI is its enormous and rapidly growing demand for electricity.

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. [6]

This expansion risks increasing electricity costs for households and industries, placing additional pressure on already strained electricity grids and diverting energy resources from other essential societal needs.

Where additional electricity demand is met through fossil fuels, the expansion of AI also risks exacerbating the climate crisis.

Even where renewable energy is used, competition for available clean electricity could impede the decarbonization of other sectors.

AI may contribute to improvements in energy efficiency and climate mitigation. However, these potential benefits cannot be assumed to compensate for the environmental consequences of unrestricted expansion.

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.

Economic and geopolitical risks

Competition over computing capacity, semiconductors, electricity, capital and technological expertise could intensify existing inequalities, industrial displacement, economic insecurity and international tensions.

AI may generate enormous economic benefits. But competition over its development and control could also contribute to conditions leading to major international conflict, potentially involving nuclear weapons.

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

The fundamental problem

No individual company or country can unilaterally ensure that other actors will refrain from developing or exercising potentially catastrophic AI capabilities.

Nor can any major power reliably protect itself against every possible consequence of unrestricted international competition.

The decisive question is whether the major powers will recognize that unilateral control cannot preserve even a chance of any protection against catastrophic AI risks.

International cooperation must therefore be based not on mutual trust, shared political values or goodwill, but on recognition of a common existential self-interest.

A global constitutional architecture

AI Governance Watch proposes the establishment of a globally binding constitutional architecture governing potentially catastrophic AI capabilities.

This architecture should establish the boundaries of permissible authority while allowing specific national, regional and technical solutions to develop incrementally within those boundaries.

Its essential elements would include:

* Binding restrictions on the development, deployment and dissemination of AI capabilities presenting catastrophic risks.
* Independent verification mechanisms that do not require complete disclosure of sensitive information or mutual trust.
* Technical safeguards preventing the unilateral exercise of potentially catastrophic authority.
* Reliable mechanisms for interrupting dangerous activities and preventing the unilateral removal of safeguards.
* An international decision-making structure in which no individual major power possesses a veto.
* Guaranteed representation of smaller countries through a quota system.
* Binding constitutional constraints on the international institutions responsible for enforcing these restrictions.

AIW has proposed 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.

At the same time, smaller countries would be guaranteed representation through a quota system.

The governance of technologies whose consequences may affect all of humanity cannot be left exclusively to the countries possessing the greatest technological, economic or military power.

Countries with limited influence over AI development may nevertheless suffer devastating consequences from its economic, environmental and geopolitical effects.

Their representation must therefore be a constitutional feature of the international governing body, rather than a matter of discretionary consultation.

This is undoubtedly a formidable political challenge.

However, the alternative may ultimately be the disappearance of any meaningful possibility of protection.

Architecture first, incremental solutions thereafter

Humanity cannot necessarily afford to follow the historical pattern of developing regulations in response to disasters that have already occurred.

A catastrophe involving advanced AI, nuclear escalation or other irreversible consequences may leave no opportunity for institutional learning.

This does not mean that every technical or regulatory problem must be solved immediately.

It means that the binding constitutional framework within which those problems can be addressed must be established before potentially catastrophic capabilities become impossible to contain.

An invitation to international action

Existing UN initiatives on AI governance provide an important opportunity to initiate discussions about such an architecture.

Governments, scientists, technology developers, trade unions, environmental organizations, civil society organizations and international institutions all have a role to play.

The objective is not to prevent technological progress or establish an unrestricted global authority.

It is to preserve the possibility that humanity can benefit from artificial intelligence without surrendering its capacity to protect itself.

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.

Humanity does not need to agree on everything.

It needs to recognize that there are certain powers no actor can safely be permitted to exercise unilaterally.

The accompanying institutional proposal and international open letter set out this architecture in greater detail and invite international consideration.

Dr Sarah Philipson
Associate Professor Emerita
AI Governance Watch (AIW)
https://ai-governance-watch.org
sarahscphilipson@mac.com

References

[1] Pachocki, Jakub. “Perplexity and Intelligence.” OpenAI, 22 November 2026.

https://openai.com/index/perplexity-and-intelligence/

[2] Rozen, Courtney. “Huang’s Return to Nvidia Brings Open-Weight AI Debate to Forefront in Washington.” Reuters, 24 November 2026.

https://www.reuters.com/world/china/huangs-return-nvidia-brings-open-weight-ai-debate-forefront-washington-2026-11-24/

[3] California Department of Justice. “It’s the Golden State, not the Wild West: Attorney General Bonta Sues OpenAI Over Cybersecurity Incidents.” 16 November 2026.

https://oag.ca.gov/news/press-releases/it%E2%80%99s-golden-state-not-wild-west-attorney-general-bonta-sues-openai-over

[4] Pratley, Nils. “Anthropic Recognises the Risks to Humanity – but Is Its Governance Really Up to the Job?” The Guardian, 16 November 2026.

https://www.theguardian.com/business/nils-pratley-on-finance/2026/nov/16/anthropic-claude-risks-to-humanity-governance-ai

[5] Milmo, Dan; Gupta, Gaya; and Down, Aisha. “Hacks, AI and Conspiracy Theories: Why Are We Having So Many Outages?” The Guardian, 22 November 2026.

https://www.theguardian.com/technology/2026/nov/22/hacks-ai-internet-outages-airlines-banks-microsoft-apple

[6] International Energy Agency (IEA). Key Questions on Energy and AI. Paris, 16 April 2026.

https://www.iea.org/reports/key-questions-on-energy-and-ai

Note on sources and evidence

Pachocki's proposal, California's lawsuit and the electricity consumption figures are confirmed by primary sources. The US safety-testing policy is attributed to Reuters, and the discussion of Anthropic's governance draws on Guardian analysis. The projected economic, environmental and geopolitical consequences remain explicitly distinguished from established events.