A Proposed Global Constitutional Architecture for Protection Against Catastrophic AI Risks

AI Governance Watch (AIW)

26 November 2026

Proposal for international consideration

Table of contents

Executive summary

1. Introduction: The nature of the challenge
2. Recent developments and their implications
3. Loss of control over AI systems
4. Deliberate misuse by human controllers
5. Systemic fragility and critical infrastructure
6. Electricity consumption, energy security and the climate crisis
7. Economic and social consequences
8. Geopolitical competition and the risk of catastrophic war
9. Why unilateral protection is insufficient
10. Existential self-interest as the basis for international cooperation
11. The fundamental constitutional principle
12. The proposed modified UN institutional structure
13. Independent verification and technical safeguards
14. Open-weight AI and irreversible dissemination
15. Constitutional limitations on international enforcement authority
16. Architecture first, incremental implementation thereafter
17. A pathway towards international negotiations
18. Outstanding questions
19. Conclusion

References

Note on sources and evidence

Executive summary

The rapid development of artificial intelligence creates unprecedented opportunities for humanity but also introduces catastrophic risks that cannot adequately be addressed through national regulation, voluntary corporate commitments or technical safeguards alone.

The dangers extend beyond the possibility of AI systems escaping human control.

They include the deliberate misuse of AI, the destabilization of critical infrastructure, increasing pressure on electricity supplies, exacerbation of the climate crisis, economic insecurity and the possibility that unrestricted technological competition could contribute to catastrophic international conflict.

The central problem is that no individual actor can ensure that other actors will refrain from developing or exercising potentially catastrophic AI capabilities.

Competitive pressures may encourage governments and corporations to develop, deploy or disseminate increasingly powerful systems despite recognizing the dangers involved.

AI Governance Watch proposes a globally binding constitutional architecture that establishes enforceable boundaries on potentially catastrophic authority.

The architecture would constrain AI systems, the actors controlling them and the international institutions responsible for enforcing those constraints.

A proposed modified UN-based institutional structure would include a 14-member council, with decisions requiring a qualified majority of nine members, without individual or collective great-power vetoes and with guaranteed representation of smaller countries through a quota system.
The global constitutional framework would establish the boundaries of permissible authority, while allowing national, regional and technical solutions to develop incrementally within those boundaries.

Its political foundation would be the recognition of common existential self-interest rather than mutual trust.
Its objective would not be to guarantee safety, but to preserve a chance of any protection against catastrophic AI risks.

Chapter 1. Introduction: The nature of the challenge

Artificial intelligence differs from many previous technological developments in the potential speed, scale and diversity of its consequences.

Increasingly capable AI systems may contribute to scientific discoveries, economic prosperity and improved human welfare. They may also enable actions whose consequences extend beyond national borders and cannot subsequently be reversed.

The challenge encompasses several interconnected categories of risk:

- Loss of control over increasingly autonomous AI systems.
- Deliberate misuse of AI by human controllers.
- Destabilization of critical infrastructure.
- Increasing electricity demand and exacerbation of the climate crisis.
- Economic and social destabilization.
- Geopolitical competition and potentially catastrophic military conflict.

These risks differ in their mechanisms, probabilities and potential consequences. They should not be treated as equally established or inevitable.

Nevertheless, they share an important characteristic: decisions taken by individual governments, corporations or other actors may have consequences extending far beyond their own jurisdictions. The governance challenge is therefore inherently international.

Chapter 2. Recent developments and their implications

Several developments during 2026 illustrate both growing recognition of the problem and the limitations of existing approaches.

2.1 Shared limits on technological advancement

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 represents an important movement towards concrete constraints rather than general commitments to responsible development.

It recognizes that individual developers cannot easily exercise restraint when their competitors continue to advance.

However, the proposal does not itself establish an internationally binding agreement or an enforcement mechanism.

Furthermore, measures of technological advancement must not automatically be equated with measures of catastrophic capability.

The challenge is to establish independently verifiable limits that remain effective when individual actors have incentives to exceed them.

2.2 The open-weight AI controversy

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]

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.

Open-weight models offer important benefits, but their dissemination can also create risks that cannot readily be reversed.

2.3 Corporate governance and external accountability

California's lawsuit against OpenAI and the public debate surrounding Anthropic's governance arrangements illustrate different dimensions of the accountability problem. [3, 4]

The allegations against OpenAI remain subject to judicial determination.

Anthropic's institutional arrangements represent a serious effort to incorporate long-term societal interests into corporate decision-making, but their effectiveness under conditions of fundamental conflict remains uncertain.

Corporate safety commitments are valuable. They cannot, however, substitute for independently enforceable constraints on potentially catastrophic authority.

2.4 Systemic fragility

Recent disruptions to transportation, communications, financial services and other essential systems have demonstrated the vulnerability of highly interconnected infrastructure. [5]

These incidents should not be attributed to AI without supporting evidence.
Nevertheless, increasing AI capabilities may interact with existing institutional and technological weaknesses in ways that amplify systemic risks.

2.5 Rapidly increasing electricity demand

The International Energy Agency reports that electricity consumption by AI-focused data centres increased by 50% in 2025 and is projected to triple between 2025 and 2030. [6]

The expansion of AI therefore raises important questions not only about technological safety, but also about energy security, electricity costs and the climate crisis.

Taken together, these developments reinforce the need for an international governance architecture addressing both direct and indirect catastrophic risks.

Chapter 3. Loss of control over AI systems

Increasingly autonomous AI systems may acquire capabilities that exceed the effectiveness of existing safeguards.

They may pursue assigned objectives in unexpected ways, circumvent restrictions or interact with external systems beyond their authorized boundaries.

Improved alignment, interpretability, monitoring and independent safety evaluations are therefore essential.

However, their effectiveness cannot be assumed, particularly as AI capabilities continue to advance. Technical progress in understanding AI systems should not automatically be interpreted as evidence that their behaviour can be reliably controlled under all relevant circumstances.

Nor should the absence of observed catastrophic behaviour be treated as proof that catastrophic behaviour is impossible.

The development and deployment of increasingly capable AI systems must therefore be accompanied by independently verifiable safeguards.

These safeguards should, wherever technically feasible, operate independently of the AI system’s own willingness to comply.

Chapter 4. Deliberate misuse by human controllers

Even an AI system that faithfully follows human instructions may perform catastrophic actions if instructed to do so.

The problem is therefore not merely how to align AI with human intentions, but how to constrain the authority of those whose intentions AI is designed to implement.

An AI system may be technically aligned and nevertheless be employed for destructive purposes.

Neither governments nor private corporations can automatically be regarded as reliable guardians of humanity’s collective interests.
Their decisions may be influenced by commercial competition, political objectives, military considerations or the pursuit of strategic advantage.

An adequate governance architecture must address the exercise of dangerous authority, irrespective of whether that authority originates in an autonomous AI system or in its human controllers.
This distinction has fundamental institutional implications.

The actors responsible for developing and controlling AI systems cannot possess unrestricted authority to determine the safeguards that constrain their own actions.

Chapter 5. Systemic fragility and critical infrastructure

Modern societies depend on increasingly interconnected technological systems.

Electricity supplies, transportation, communications, financial services, healthcare and public administration rely on infrastructures in which failures can propagate across institutional and national boundaries.

Recent outages have demonstrated how apparently limited technical failures can produce extensive societal disruption.

The increasing use of AI may introduce additional vulnerabilities into these systems.

AI-assisted discovery of software vulnerabilities can improve cybersecurity. However, the speed and volume of such discoveries may also place considerable pressure on organizations responsible for repairing and maintaining essential infrastructure.

Rapidly implemented repairs may themselves introduce new failures.

Furthermore, increasingly capable AI systems may enhance the ability of malicious actors to exploit vulnerabilities.

The resulting risks cannot be addressed solely through improving the intrinsic safety of individual AI models.
Governance must also consider the environments in which AI capabilities are exercised, including the resilience of critical infrastructure, the concentration of technological dependencies and the adequacy of contingency arrangements.

Chapter 6. Electricity consumption, energy security 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 has several important implications.

6.1 Electricity costs and energy security

The growing demand for electricity from AI and data centres risks increasing energy costs for households, public services and industries.

The magnitude of these effects will vary between countries and regions, depending on electricity generation capacity, grid infrastructure, regulatory arrangements and the allocation of investment costs.

Nevertheless, rapid expansion can place additional pressure on already strained electricity grids and require substantial investment in generation, transmission and distribution infrastructure.

An important question is who will bear the costs of these investments.

Households and industries that do not directly benefit from the expansion of AI may nevertheless face higher electricity prices or reduced access to affordable energy.

6.2 Exacerbation of the climate crisis

Where additional electricity demand is met through fossil fuels, the expansion of AI risks increasing greenhouse gas emissions and delaying decarbonization.
Even where renewable energy is used, competition for available clean electricity could impede the decarbonization of other sectors.

The environmental consequences of AI development must therefore be assessed in relation to the broader energy transition, rather than solely in terms of the electricity sources used by individual data centres.

AI may contribute to improvements in energy efficiency, climate modelling and 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.

6.3 International competition and energy allocation

Electricity is not merely an input into AI development. It is an essential resource for the functioning of modern societies.

Competition over electricity supplies could intensify inequalities between countries and between different sectors of national economies.

Countries with limited capacity to participate in the AI race may nevertheless be affected by its environmental consequences, rising energy costs and changes in international resource allocation. These consequences reinforce the importance of ensuring that smaller countries participate in international AI governance.

6.4 Governance implications

International AI governance should incorporate systematic assessment of the energy and climate consequences of AI development.

This should include consideration of electricity consumption, greenhouse gas emissions, infrastructure requirements and the distribution of associated costs.
International cooperation should seek to prevent competitive pressures from undermining climate commitments or energy security.

However, this does not imply that an international AI governing body should possess unrestricted authority over national energy policies or global resource allocation.

Any binding responsibilities in these areas must be clearly defined and constitutionally constrained.

Chapter 7. Economic and social consequences

The AI race may intensify competition over semiconductors, computing infrastructure, energy, capital, industrial production and technological expertise.

These developments could contribute to economic insecurity, industrial displacement, strategic dependencies, protectionism and increasing inequalities.

The concentration of AI capabilities and infrastructure may also concentrate economic and political power.

Countries, industries and workers unable to participate on comparable terms may experience consequences over which they exercise limited influence.

AI could simultaneously produce substantial economic benefits, improve productivity and create new opportunities.

The distribution of these benefits and costs, however, cannot be assumed to be equitable or socially sustainable.

Trade unions, civil society organizations and representatives of affected communities should participate in international discussions concerning AI’s consequences for employment, industrial development, economic security and social stability.

The governance of catastrophic AI risks cannot be entirely separated from the economic and social conditions created by technological competition.
Nevertheless, the authority to address catastrophic risks must be distinguished from general authority over the allocation of economic resources.

Recognition of these dangers must not become a justification for unrestricted international control over economic activity.

Chapter 8. Geopolitical competition and the risk of catastrophic war

Competition over artificial intelligence is increasingly intertwined with national security, economic competitiveness and geopolitical influence.

Governments may perceive technological superiority as essential to their national security.

Restrictions on AI development may consequently be interpreted as threats to strategic competitiveness, particularly when competing powers are not subject to equivalent restrictions.

The resulting competition may intensify existing international tensions.

There are at least two distinct pathways through which the AI race could contribute to catastrophic military conflict.

8.1 Asymmetric recognition of catastrophic capabilities

If one major power becomes convinced that another is approaching an irreversible and potentially catastrophic technological advantage, it may consider preventive military action.

This danger could be particularly acute if different powers reach fundamentally different conclusions about the urgency or magnitude of the threat.

Under certain circumstances, the resulting conflict could escalate into a nuclear confrontation.

8.2 Economic and resource competition

Competition over energy, semiconductors, industrial production, capital and strategic resources could contribute to economic destabilization, protectionism and international conflict.
Such developments could intensify existing geopolitical rivalries and contribute to conditions leading to a third world war.

Neither pathway is inevitable.

However, both demonstrate that humanity could face catastrophic consequences from the AI race even without AI itself becoming uncontrollable.

International AI governance must therefore consider the risks generated by competition over AI, as well as the risks generated by AI capabilities themselves.

Chapter 9. Why unilateral protection is insufficient

No major power can independently determine the actions of all other governments, corporations or non-state actors.

A country may impose stringent domestic restrictions while remaining exposed to dangerous capabilities developed elsewhere.

Conversely, a country may seek protection through technological superiority, only to encourage competing powers to accelerate their own development.

This creates a security dilemma in which individually rational decisions may produce collectively catastrophic consequences.

The global dissemination of increasingly capable AI systems further complicates unilateral protection.

Potentially catastrophic capabilities may become accessible beyond the jurisdictions in which they were originally developed.

Neither national regulation nor technological superiority can reliably eliminate these vulnerabilities. The question is not simply whether international cooperation would provide better protection than unilateral action.
It is whether unilateral action can preserve even a chance of any protection.

Chapter 10. Existential self-interest as the basis for international cooperation

The establishment of effective international governance does not require governments to trust one another.

Nor does it require agreement on political systems, economic models, ideological principles or broader geopolitical objectives.

The decisive political condition may instead be recognition of a common existential self-interest. The major powers must recognize that unilateral control cannot preserve even a chance of any protection against catastrophic AI risks.

This recognition would not eliminate international rivalry.

It could, however, establish a basis for agreement on certain limits that all participants have an existential interest in maintaining.

Such an agreement would require major powers to accept constraints on their own authority, including decisions with which they fundamentally disagree.

It would also require them to accept that the interests of smaller countries cannot be excluded from decisions whose consequences may affect all of humanity.

The political feasibility of such an arrangement remains uncertain.

However, its difficulty must be assessed against the potentially catastrophic consequences of failing to establish any effective international constraints.

Chapter 11. The fundamental constitutional principle

The proposed architecture is based on a fundamental principle:

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

This principle applies to governments, corporations, AI systems and international institutions.

The objective is not to centralize all authority over artificial intelligence.

It is to establish binding boundaries beyond which potentially catastrophic authority cannot legitimately be exercised.

The term constitutional refers here to the establishment of fundamental rules governing the distribution, limitation and exercise of authority.

These rules must bind those responsible for enforcing them as well as those subject to enforcement.

The architecture must therefore address two problems simultaneously:

Preventing the unrestricted exercise of potentially catastrophic authority by individual actors.

Preventing the international enforcement institution from acquiring unrestricted authority itself.

Resolving only the first problem could simply transfer catastrophic authority from national or corporate actors to a global institution.

An effective architecture must address both.

Chapter 12. The proposed modified UN institutional structure

AIW proposes the establishment of a modified UN-based institutional structure with responsibility for governing potentially catastrophic AI capabilities.

12.1 Composition and decision-making

A central element would be a council consisting of 14 members.

Decisions within its defined mandate would require a qualified majority of nine of the 14 members. No individual member would possess a veto.

In particular, the United States, China and Russia would have to accept the possibility that binding decisions could be adopted against their wishes.

Nor would these three powers collectively possess a veto.

12.2 Guaranteed representation of smaller countries

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. Smaller countries must be guaranteed representation through a quota system.

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 governing body, rather than a matter of discretionary consultation.

The quota arrangements must be incorporated into the composition of the council itself.

12.3 Constitutional legitimacy and accountability

Qualified-majority voting and guaranteed representation are necessary elements of the proposed architecture, but cannot alone guarantee legitimate or effective governance.

The council’s mandate, membership, procedures, accountability mechanisms and relationship to existing UN institutions must be constitutionally specified.

Its powers must be restricted to the purposes defined in its founding agreement.

Independent review, transparency requirements, conflict-of-interest safeguards and mechanisms for challenging abuses of authority must form part of the architecture.

12.4 Legal foundations

The appropriate legal route, including whether changes to existing UN arrangements or a separate binding international agreement would be necessary, requires further examination.
The objective is not merely to establish another international forum for discussion.

It is to establish an institution capable of adopting binding decisions within a clearly defined and constitutionally constrained mandate.

Chapter 13. Independent verification and technical safeguards

International rules will be insufficient unless accompanied by practical mechanisms capable of verifying and enforcing compliance.

The architecture should establish requirements in the following areas.

13.1 Independent verification

Compliance must be capable of independent assessment.

Verification mechanisms should protect legitimate national security interests, commercial confidentiality and sensitive technical information.

The objective is to verify compliance without requiring complete disclosure or mutual trust.

The feasibility and limitations of particular verification methods must be independently evaluated.

13.2 Measurable development and capability thresholds

Pachocki’s proposal illustrates the potential importance of establishing shared, measurable limits on technological advancement.

However, such measurements must be evaluated in relation to the actual capabilities and risks of AI systems.

Development thresholds should not be treated as reliable indicators of catastrophic risk without sufficient supporting evidence.

The architecture should provide a mechanism for establishing, reviewing and revising appropriate thresholds as scientific understanding develops.

13.3 Limits on dangerous permissions

Potentially catastrophic AI capabilities must be subject to enforceable restrictions on the actions they can perform and the resources they can access.

Such restrictions should not depend exclusively on the AI system’s willingness to comply with instructions.

Where technically feasible, safeguards should be enforced through mechanisms external to the system itself.

13.4 Distributed authorization

Actions involving potentially catastrophic consequences should not be subject to unilateral authorization.

Appropriate forms of distributed authorization should prevent individual governments, corporations, officials or AI systems from independently exercising defined catastrophic capabilities.

These mechanisms must themselves be protected against collusion, circumvention and abuse.

13.5 Interruption and containment

Systems possessing potentially catastrophic capabilities should be subject to independently assessable mechanisms for interruption, containment and restriction of unauthorized activities. The architecture must address circumstances in which these safeguards fail or are deliberately circumvented.

13.6 Protection against unilateral removal of safeguards

Safeguards cannot provide meaningful protection if the actors they constrain can remove them whenever doing so becomes commercially or strategically advantageous.

The architecture should therefore establish independently enforceable requirements governing the modification, suspension or removal of critical protections.

Chapter 14. Open-weight AI and irreversible dissemination

The international debate over open-weight AI illustrates the tension between technological accessibility, commercial competition, national security and collective safety.

Open-weight models can promote innovation, scientific research and competition while reducing dependence on dominant technology corporations.

These are substantial benefits.

However, public dissemination of models possessing sufficiently dangerous capabilities may make subsequent restrictions practically unenforceable.

Once model weights are widely distributed, they may be copied, modified and deployed beyond the effective control of their original developers.

Safeguards incorporated into the original systems may subsequently be removed or circumvented. Proprietary models create different risks, including the concentration of technological authority in a small number of companies or governments.

Neither unrestricted dissemination nor concentrated proprietary control constitutes an adequate governance solution.

The relevant criterion should be the potentially catastrophic capabilities involved, rather than whether a particular system is open-weight or proprietary.

The public dissemination of models possessing defined catastrophic capabilities should be subject to appropriate prior assessment and binding restrictions.

These restrictions should be capability-based, proportionate and designed to preserve legitimate scientific research and beneficial technological development.

Chapter 15. Constitutional limitations on international enforcement authority

The creation of a global enforcement institution could itself introduce catastrophic risks if that institution acquired unrestricted technological, political or coercive authority.

An international governance architecture must therefore avoid merely transferring the problem of unconstrained authority from national to global institutions.

Its constitutional safeguards should include:

- A narrowly defined mandate.
- Distributed decision-making authority.
- Independent oversight and review.
- Restrictions on unilateral emergency powers.
- Protection against the concentration of unrestricted technological capabilities.
- Accountability for actions exceeding constitutionally authorized powers.
- Mechanisms for challenging and correcting abuses of authority.

The institution responsible for enforcing restrictions must not possess an unrestricted capacity to exercise the very powers it is intended to constrain.

Its responsibilities concerning indirect economic, environmental and geopolitical risks must likewise be clearly defined.

The recognition that AI competition may exacerbate the climate crisis or contribute to economic destabilization must not become a justification for unrestricted global authority over energy, economic or national policies.

Chapter 16. Architecture first, incremental implementation thereafter

The proposal does not assume that every technical, legal or institutional problem can be resolved immediately.

Nor does it require uniform global regulation of every application of artificial intelligence.

The distinction is between the establishment of binding constitutional boundaries and the subsequent development of specific solutions within those boundaries.
National governments, regional institutions, scientific organizations and technology developers should retain substantial freedom to develop appropriate approaches consistent with the global framework.

Technical standards and verification methods must evolve as AI capabilities change.

However, the fundamental constraints on potentially catastrophic authority must not depend solely on voluntary participation or unilateral decisions.

Historical approaches to technological regulation have frequently involved learning from accidents and disasters.

Where the potential consequences are irreversible, this approach may no longer be adequate.
A catastrophe involving advanced AI, nuclear escalation or other irreversible consequences may leave no opportunity for institutional learning.

The constitutional architecture must therefore precede the full realization of the catastrophic risks it is intended to address.

Chapter 17. A pathway towards international negotiations

Existing UN initiatives on artificial intelligence provide a potential institutional starting point.

AIW proposes that forthcoming international discussions consider establishing a dedicated process to examine the feasibility of a binding constitutional architecture.

Such a process could include:

1. An international assessment of potentially catastrophic AI capabilities and indirect catastrophic pathways.
2. Examination of the legal and institutional options for establishing binding international constraints.

3. Independent assessment of technically feasible verification and enforcement mechanisms.
4. Negotiations concerning decision-making procedures, including the proposed qualified majority of nine members in a 14-member council.
5. Constitutional specification of guaranteed quota-based representation for smaller countries.
6. Development of constitutional safeguards limiting the authority of international enforcement institutions.
7. Assessment of the electricity, climate, economic and geopolitical consequences of unrestricted AI competition.
8. Consultation with scientific institutions, technology developers, trade unions, environmental organizations and civil society.
9. Negotiation of an implementation framework incorporating evolving technical standards and national and regional solutions.

These preparatory activities could begin immediately, without presupposing agreement on every element of the final architecture.

The objective should be to move beyond general declarations of shared concern towards consideration of concrete, binding institutional arrangements.

Chapter 18. Outstanding questions

The proposal is intended as a basis for international consideration rather than a claim that all institutional and technical problems have been resolved.

Important questions include:

How should potentially catastrophic capabilities be defined and assessed?

Which verification mechanisms can operate effectively without requiring unacceptable disclosure?

How should measurable development thresholds be related to catastrophic capabilities?

How should the proposed council’s membership and quota-based representation be implemented?

What enforcement mechanisms would be both effective and constitutionally constrained?

How should persistent non-compliance by major powers be addressed?

- How can existing dangerous capabilities and previously disseminated models be brought within an effective governance framework?
- How can international enforcement avoid generating new geopolitical conflicts?
- How can the energy and climate consequences of AI development be addressed without establishing excessive international authority over energy policies?
- How can indirect economic and geopolitical risks be addressed without establishing excessive international economic authority?
- What institutional mechanisms would ensure meaningful participation by trade unions, environmental organizations, scientific institutions and civil society?

These questions require serious interdisciplinary investigation.

Their difficulty does not diminish the importance of addressing them before potentially catastrophic developments become irreversible.

Chapter 19. Conclusion

Humanity faces a challenge that cannot be resolved solely through improved AI alignment, responsible corporate leadership, technological superiority or fragmented national regulation.

The dangers extend beyond the possibility that AI systems might escape human control. They include deliberate misuse, systemic fragility, increasing pressure on energy supplies, exacerbation of the climate crisis, economic destabilization and potentially catastrophic international conflict.

The fundamental institutional issue is the distribution and limitation of potentially catastrophic authority.

An effective governance architecture must constrain AI systems, those controlling them and the institutions established to enforce those constraints.

It must also recognize that countries with limited technological, economic or military power may nevertheless be profoundly affected by the consequences of AI development.

Their representation must therefore be constitutionally guaranteed.

The proposed architecture does not assume that governments will trust one another or that corporations will consistently prioritize collective interests.

It rests on the possibility that competing actors will recognize a common existential self-interest in preventing the unrestricted exercise of catastrophic capabilities.

The objective is not to guarantee humanity's survival.

It is to preserve a chance of any protection.

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

Humanity must establish enforceable limits on catastrophic authority before the capacity to establish such limits is lost.

References

[1] Pachocki, Jakub. "Perplexity and Intelligence." OpenAI, 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, April 2026. https://www.iea.org/reports/key-questions-on-energy-and-ai

Note on sources and evidence

This proposal distinguishes between documented developments, reported policy decisions, scientific projections and AIW’s analysis of future risks.

Pachocki’s proposal, California’s lawsuit and the electricity consumption figures are drawn from primary sources. The US safety-testing policy is attributed to Reuters, and the discussion of Anthropic’s governance draws on Guardian analysis.

The allegations made in the California lawsuit have not been treated as judicially established facts.

The projected economic, environmental and geopolitical consequences of unrestricted AI competition represent risk assessments rather than predictions of inevitable outcomes.

The proposed constitutional architecture constitutes AIW’s institutional response to these risks.

AI Governance Watch (AIW)

Dr Sarah Philipson

Associate Professor Emerita

ai-governance-watch.org

ai-governance-watch.com

Submitted as a proposal for international discussion and further institutional, legal and technical development.