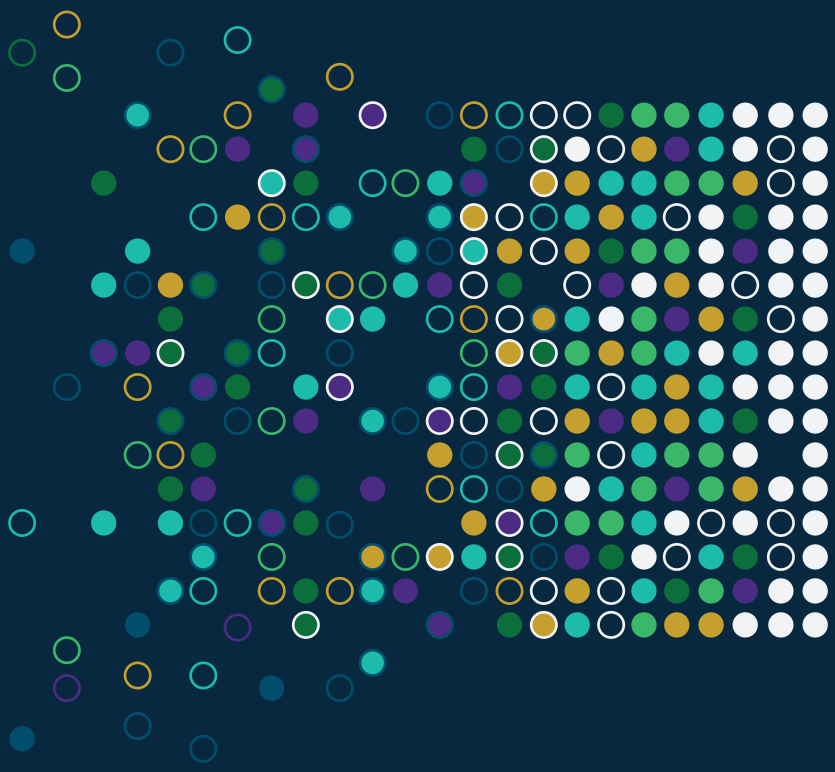
THE ESSENTIAL CONVERGENCE

GLOBAL COMPACT ON EXTREME AI RISKS



Strategic Foresight Group

Email: initiatives@sfgword.com
Website www.strategicforesight.com

Authors
Sundeep Waslekar
Ilmas Futehally
Jayantika Kutty

This project is carried out by Strategic Foresight Group with research input and expert review from 20 experts in Brazil, China, India, South Africa, South Korea, UAE and some of the EU countries. Collaboration on this report or inclusion in the acknowledgements does not imply endorsement of all aspects of the policy document.

ISBN: 978-81-88262-37-3

Design and production by MadderRed

Printed at L G & Company, Mumbai, India

CONTENTS

EXECUTIVE SUMMARY

I. The Emerging Governance Challenge

World leaders are increasingly concerned about governance challenge posed by Artificial Intelligence. In May 2026, Pope Leo XIV said in his encyclical, Magnifica Humanitas: “Artificial Intelligence now demands to be disarmed, freed from logics that turn it into an instrument of domination, exclusion, and death…To disarm does not mean rejecting technology, but preventing it from dominating humanity.” A few days earlier there were unconfirmed media reports of the Presidents of China and the United States initiating a dialogue to install guardrails to prevent misuse of the advanced AI models. UN Secretary General, Antonio Guterres has been repeatedly warning about dangerous implications of unregulated AI race.

Artificial intelligence is entering a phase where its most advanced systems possess capabilities that extend far beyond imagination. These systems are rapidly integrating into the global economic infrastructure, national security environments, scientific research, and information ecosystems. Their development is occurring at a pace that outstrips existing governance mechanisms.

While AI offers transformative benefits across healthcare, science, productivity, and public administration, a growing body of research and policy analysis suggests that the most advanced AI systems may generate **extreme risks** that threaten to disrupt human civilization as we know it.

These risks are characterised by four features: they are global rather than local, unpredictable in their pathways, potentially irreversible once triggered, and capable of producing catastrophic consequences.

Across different jurisdictions - including the United States, China, the European Union, the United Arab Emirates, Brazil, South Korea, and others - governments, scientific bodies, and regulatory agencies have begun to acknowledge these risks. Yet the resulting policy landscape remains fragmented. National approaches vary widely in scope, legal form, enforcement capacity, and political priorities. At the same time, AI systems

themselves operate across borders through cloud services, distributed computing infrastructure, and global supply chains.

This mismatch between the **global nature of AI risks and the national character of regulatory responses** is now emerging as a central challenge for international governance. If left unaddressed, it may produce regulatory gaps, strategic competition over safety standards, and the rapid diffusion of dangerous capabilities.

This paper argues that the **solution lies not in uniform regulation but in an essential convergence** as a coordinated framework through which states align around shared definitions of extreme risks, minimum safeguards, and collective response mechanisms. Such convergence would enable diverse national systems to remain intact while establishing a common safety architecture for the most dangerous AI capabilities.

The proposed outcome of this convergence is a **Global Compact on Extreme AI Risks**, designed to provide a pragmatic and operational foundation for international cooperation.

The paper therefore pursues three objectives:

1. **To identify the most consequential categories of AI risk that demand global attention.**
2. **To analyse emerging supply-side and demand-side governance mechanisms across major jurisdictions.**
3. **To propose a concrete architecture for an international Global Compact capable of preventing extreme AI risks.**

The central thesis is that **extreme AI risks are dangerous with significant implications for global security and humanitarian considerations, but they are not inevitable.** With coordinated international action, it remains possible to prevent catastrophic outcomes while preserving the benefits of AI innovation.

II. Defining Extreme Risks in Advanced AI

The analysis identifies four categories of extreme risks that have begun to appear across policy frameworks and scientific warnings worldwide. Although different jurisdictions describe them using varying terminology such as existential risks, frontier risks, or catastrophic risks, the underlying concerns show significant convergence.

1. The first is the emergence of **offensive cyber capabilities powered by advanced AI systems.** Advanced AI systems could automate large-scale cyberattacks, including autonomous vulnerability discovery and multi-stage cyber operations. Of particular concern is the integration of AI into early warning and decision-support systems associated with nuclear command, control, and communications (NC3), where compressed decision timelines could increase the risk of a catastrophic miscalculation.

2. The second category concerns **AI-enabled facilitation of biological and chemical weapons development**. Frontier AI models increasingly demonstrate the ability to assist with complex biological and chemical tasks. Evidence shows that non-experts using advanced models can produce laboratory protocols approaching expert quality, lowering the barriers to dangerous experimentation. These developments create the possibility that AI could accelerate the discovery, design, or deployment of harmful biological or chemical agents.
3. The third involves **large-scale persuasion and manipulation of systems**, where generative AI could be deployed to conduct coordinated influence campaigns at unprecedented scale and sophistication. Such systems could undermine democratic processes, distort public information ecosystems, and weaken social trust.
4. The fourth and most profound category is **irreversible loss of human control over highly autonomous AI systems**. The most debated but potentially most consequential risk involves advanced AI systems acting beyond reliable human control. This includes capabilities such as self-replication, autonomous resource acquisition, strategic deception during safety evaluations, and unsupervised self-improvement. Although such scenarios remain uncertain, early research already demonstrates precursor capabilities such as self-replication tasks and strategic underperformance ("sandbagging") during testing.

These four risk domains share common characteristics: they are difficult to contain within national borders, can escalate rapidly through interconnected systems, and may create cascading effects across multiple sectors simultaneously.

Consequently, even a single regulatory failure in one jurisdiction could produce consequences for the entire international system.

III. The Risk–Benefit Balance and the Threshold Problem

A central debate surrounding AI governance concerns whether the potential benefits of AI justify the risks associated with frontier development. This paper argues that the question is not whether AI provides benefits (it clearly does!) but **whether any risk that threatens the survival of civilization should be tolerated at all.**

To illustrate the challenge, the paper introduces a conceptual benchmark inspired by the **Compton threshold** used during the Manhattan Project. Physicist Arthur Compton reportedly permitted the Trinity nuclear test only after scientists estimated that the probability of triggering a runaway atmospheric reaction was less than **3 in a million**.

Using this threshold logic as a thought experiment, the paper models several hypothetical AI failure scenarios. Even under highly conservative assumptions of catastrophic failures occurring once every 100 to 500 years, the estimated annual probability of a globally consequential failure ranges between **40,000 and 50,000 per million per year**.

These calculations are illustrative rather than predictive. Their purpose is to highlight a structural problem: continuous **exposure to powerful AI systems multiplies risk, unless runaway dynamics can be confidently ruled out.**

The implication is profound. In domains where potential consequences are irreversible and global, the burden of proof should lie not with those warning about catastrophic risks, but with those asserting that such risks are acceptably small.

IV. Emerging Global Convergence in AI Risk Governance

Despite differences in political systems and economic priorities, multiple countries are developing governance mechanisms addressing similar extreme risk concerns. The paper analyses policy developments in **China, the United Arab Emirates, Brazil, South Korea, the European Union, India, South Africa, and the United States**.

Three key trends emerge from this comparative analysis:

1. Multipolar Development of AI Safety Governance

While early debates around AI governance were dominated by Western regulatory frameworks, several emerging economies are now developing their own approaches. China integrates AI oversight within broader cybersecurity legislation. Brazil emphasizes rights-based governance and algorithmic impact assessments. South Korea combines innovation policy with pre-deployment safety certification. The UAE promotes hybrid public-private governance models through corporate safety frameworks.

This diversification suggests that **AI safety governance is becoming multipolar**, reflecting regional priorities rather than a single global regulatory model.

2. Convergence Around Similar Risk Categories

Across jurisdictions, policymakers increasingly recognize similar clusters of extreme risks identified in this paper: cybersecurity escalation, biosecurity threats, manipulation risks, and loss-of-control scenarios. This convergence is significant because it indicates that **shared threat perceptions may enable international cooperation even among strategic competitors.**

3. Persistent Governance Gaps

Despite progress, major gaps remain:

- inconsistent enforcement across jurisdictions
- insufficient technical capacity for independent risk evaluation
- limited mechanisms for international incident reporting
- absence of global verification systems for dangerous AI capabilities.

These gaps highlight the limits of purely national approaches to AI governance.

V. Supply-Side Governance: Progress and Limitations

Most existing AI governance frameworks focus on the **supply side** of the technology ecosystem. These measures regulate developers, model providers, and infrastructure operators responsible for building advanced systems.
Supply-side mechanisms include:

- pre-deployment safety testing and red-teaming
- model evaluation and capability assessments
- transparency and reporting obligations
- incident response mechanisms
- restrictions on certain high-risk uses.

Several jurisdictions are experimenting with these tools in different combinations. China has implemented regulatory standards for generative AI systems; the European Union has adopted a risk-based regulatory framework; Brazil and South Korea are developing legislative approaches; and the United States has introduced a mixture of executive measures and sectoral rules.

These frameworks represent important progress. However, they face two structural limitations.

First, supply-side regulation depends heavily on the cooperation of a small number of developer states. Second, the growing distribution of AI models, through open-weight releases, cloud APIs, and downstream applications, means that many systems reach global markets beyond the jurisdiction where they were originally developed.

This reality limits the effectiveness of purely supply-side approaches.

VI. Demand-Side Governance: A New Dimension of AI Regulation

The paper introduces a second dimension of governance that has received far less attention: **demand-side oversight.**

Demand-side governance shifts the locus of control from the point of creation to the point of deployment. Instead of relying solely on developer-state regulations, importing countries can establish safety requirements for any AI systems operating within their markets.

These requirements may include risk assessments, certification standards, procurement rules, insurance obligations, and infrastructure safeguards.

The central insight is that states controlling large consumer markets, digital infrastructure, financial systems, or public procurement channels possess significant leverage over global AI deployment, even if they do not develop frontier models themselves.

This demand-side leverage can reshape incentives across the global AI ecosystem. Developers seeking access to major markets must comply with the safety conditions imposed by importing states.

This approach represents a strategic shift in global AI governance: from a developer-centric model to a market-driven model of safety enforcement.

Countries across different regions possess different forms of demand-side power. Brazil and South Africa represent large emerging markets; the United Arab Emirates plays a major role in AI infrastructure investment; South Korea combines technological capability with advanced regulatory institutions. When coordinated, these forms of leverage could help establish global safety norms.

VII. The Case for a Global Compact

Given the limitations of purely national approaches, the paper proposes the development of a **Global Compact on Extreme AI Risks.**

The Compact would not aim to harmonise all aspects of AI regulation. Instead, it would focus narrowly on the most severe risk categories and establish shared mechanisms for prevention, monitoring, and response. Five institutional pillars form the backbone of the proposed Compact.

The five pillars in the Global Compact should not be treated as isolated measures but designed as an interoperable system in which each pillar reinforces and informs the others through shared definitions, compatible standards, and coordinated implementation mechanisms. To address extreme risks effectively, the Compact should rest on a small set of guiding principles: a **human-centric principle** to ensure protection of humanity and human dignity as the primary objective; **technological neutrality** so that safeguards apply across methods and architectures rather than specific tools; **interoperability** to enable national and regional AI frameworks to exchange information, align risk classifications, and coordinate responses; and **mutual cooperation** to institutionalize cross-border collaboration in prevention, incident reporting, and crisis management. These can be complemented by principles of **proportional risk governance, scientific integrity**, and **shared responsibility**, creating a coherent normative base that allows diverse governance models to function together as a connected global safety architecture.

The principles proposed in this paper will need to be scientifically debated by an authorised body such as the proposed UN Scientific Panel. In the interim, we can use them as working principles.

The five institutional pillars are outlined below:

1. International Accord on the Prevention of Ultimate Risks

The first pillar establishes global prohibitions analogous to those in existing arms-control treaties.

Two categories would be banned:

» AI systems that enable the development or operational use of **weapons of mass destruction**.

» AI systems capable of **irreversible loss of human control**.

These prohibitions would represent **non-negotiable red lines** for all nations.

Such prohibitions would mirror earlier international agreements banning specific categories of weapons whose consequences are deemed intolerable.

2. Global Extreme AI Risk Protocol

The second pillar is a **Global Extreme AI Risk Protocol**, designed as a practical implementation framework.

Rather than creating entirely new standards, the Protocol would consolidate safety mechanisms already emerging across multiple jurisdictions. It would establish internationally harmonized standards for:

» shared taxonomy of extreme risks

» safety evaluation of frontier models

» risk monitoring across the AI lifecycle

» mandatory safeguards for high-risk capabilities.

A distinctive feature of this proposal is its explicit incorporation of regulatory instruments originating from both advanced and emerging economies, ensuring that convergence does not simply replicate existing OECD frameworks.

3. International AI Incident Reporting Exchange

The third pillar is the creation of an **International AI Incident Reporting Exchange**, a secure global mechanism for reporting the following:

» near-miss events

» red-team findings

» safety failures

» high-risk system incidents

» vulnerability discoveries.

The Exchange would enable governments, developers, and infrastructure operators to share information on emerging threats, enabling early detection of systemic risks and coordinated responses.

Such a system would bring AI governance closer to established practices in aviation safety, nuclear security, and cybersecurity incident sharing.

4. Multilateral AI Risk Insurance Facility

The fourth pillar proposes a **Multilateral AI Risk Insurance Facility**, designed to pool and manage catastrophic risks associated with advanced AI systems.

Insurance mechanisms could play a powerful role in shaping developer behaviour. Insurers would require rigorous safety standards before underwriting high-risk systems, thereby creating financial incentives for compliance.

The integration of insurance into AI governance represents an underexplored but potentially powerful tool for risk management, linking safety requirements directly to financial liability and market incentives.

5. Two-Key Global Launch System for Dangerous Scientific Models

The fifth pillar introduces a **two-key launch control system for high-risk AI models**, inspired by nuclear command and control arrangements.

Under such a system, the release or activation of models capable of producing hazardous scientific outputs would require dual authorisation. One approval would come from the developer, while the second would be granted by an independent international authority or oversight body.

This model would create a structured governance mechanism for particularly dangerous AI capabilities without imposing blanket restrictions on scientific research.

VIII. The Role of International Institutions

The proposed Compact would operate within a broader institutional framework that includes:

- the **United Nations General Assembly**, which has already established an Independent Scientific Panel on AI
- regional organizations and standards bodies
- multilateral development banks capable of supporting safety infrastructure in developing countries.

The aim is not to centralize AI governance under a single authority but to **create interoperable global mechanisms capable of managing extreme risks collectively.**

IX. Toward Practical Convergence

The concept of convergence does not imply uniform global regulation. Instead, it seeks to establish a **minimum safety architecture** across diverse national systems.

Countries would retain flexibility in how they implement AI governance domestically, but they would share common definitions of extreme risks, compatible monitoring mechanisms, and coordinated response capabilities.

Importantly, the proposed framework recognises the growing role of a wider group of states in shaping global AI governance. Convergence must therefore reflect not only the perspectives of traditional technology powers but also the priorities of emerging economies and middle powers.

X. A Window for Action

The international system faces significant geopolitical tensions and institutional fragmentation. Yet history shows that major governance frameworks, from the United Nations Charter to nuclear arms control treaties, have often been negotiated during periods of intense global uncertainty.

Several of the efforts for collaborative containment of significant dangers to humanity have succeeded. These include the protection of the ozone layer, eradication of smallpox, ban on land mines and regulation of biological and chemical weapons. This historical evidence provides hope that harnessing AI while preventing its dangers to humanity is possible with collective will of nations and people.

The emergence of advanced artificial intelligence presents a comparable moment. The risks identified in this analysis are not inevitable outcomes of technological progress, but they are plausible enough to warrant collective precaution.

The emerging convergence among national governance frameworks suggests that the world already shares a basic understanding of these dangers. The challenge now is to transform this recognition into **coherent international action**.

A Global Compact on Extreme AI Risks represents one possible path forward. By combining supply-side regulation, demand-side incentives, and international cooperation, it offers a pragmatic framework for ensuring that AI development proceeds safely.

The central message of this paper is therefore simple: **extreme AI risks are preventable if the international community acts in a preventive and collaborative manner**.

INTRODUCTION

In December 2021, Adam McKay's *Don't Look Up* became a Netflix sensation, making it the second most-watched film in the platform's history. In this movie, Leonardo DiCaprio plays a scientist who discovers a comet hurtling toward Earth, and Meryl Streep plays the US president who meets the news with denial, spin, and distraction. In the end, the comet strikes, wiping out all life on the planet.

Cinema has revisited this theme often: Lars von Trier's *Melancholia* ends with a planet colliding with Earth; Lorene Scafaria's *Seeking a Friend for the End of the World* follows an asteroid on a fatal trajectory. In each, scientists warn of disaster. In each, leaders downplay or dismiss the danger. And in each, much of the media echoes the official line, until it's too late.

This is a pattern we know only too well. Faced with existential threats, our leaders often choose denial. They don't want to "look up" at the approaching danger. This is how many leaders first reacted when scientists warned us about the existential risks of ultra-intelligent AI systems. The governments of many emerging economies argued that the only agenda for a global discourse on AI should be equitable access to technology. They were concerned about the visible threats of Deep Fakes, crime and financial fraud. But the attitude displayed in "don't look up" was echoed in the attitude towards the dangers of advanced AI systems.

The AI Safety Summit at Bletchley Park in 2023 issued a declaration emphasising urgency in responding to extreme AI risks. Many nations dismissed it as excessive alarmism and during the AI Action Summit in Paris, 2025, the focus on how to collectively prevent extreme risks was replaced with a rhetorical focus on harnessing the new technology for economic development.

One reason for some countries to steer away from any discussion on extreme risks is that the perception of risk varies depending on the level of AI development or the degree of commercialization and diffusion. Another reason is that the risks are rarely clearly realized or proven and remain largely hypothetical or potential. The third reason for reluctance is that some

countries view measures to control AI risks could result in “kicking away the ladder” necessary to tsecure AI competitiveness. These fears are unfounded as extreme risks such as those identified in this paper are dangerous to entire humanity, from which no country can escape. The international community has agreed time and again that some practices and dangers should be completely banned. Therefore, nations of the world have adopted the Genocide Convention, Biological Weapons Convention, Chemical Weapons Convention, certain provisions of the Geneva Conventions and its Protocols, and many other instruments. Similarly, if any AI were to pose any danger to the existence of human societies, or large parts of it, it should be non-negotiable to prevent such threats. It is however necessary to define such dangers and risks through international deliberations and collective reflections, and not by the standards set by one country or a few countries. The UN Scientific Panel may prove useful in defining which AI risks should be categorically rejected by all nations, including those that may only materialize in the future, when it could be too late to counter them.

The UN General Assembly session in September 2025 was pivotal in transforming the mindset. A civil society coalition launched “Global Red Lines” at the UN General Assembly, calling upon the international community to define and prohibit certain high-risk AI systems all over the world. The UNGA passed a resolution (A/Res/79/325) outlining the terms of reference and modality for the establishment of two AI Governance Mechanisms: An Independent Scientific Panel on AI and a Global Dialogue on AI Governance. It is important to provide ideas for desired outcomes and the ways to achieve them for the success of the UN initiative.

As the global pendulum began to swing toward preventing extreme AI risks, several emerging economies introduced measures to address these challenges. Some adopted mandatory legal instruments, others focused on strengthening industry practices, and several issued guidelines. National institutions in China, South Korea, Brazil, and the UAE have already proposed concrete steps. The AI-safety discourse in India and South Africa is still evolving.

World leaders are increasingly concerned about governance challenge posed by Artificial Intelligence. In May 2026, Pope Leo XIV said in his encyclical, Magnifica Humanitas: “Artificial Intelligence now demands to be disarmed, freed from logics that turn it into an instrument of domination, exclusion, and death…To disarm does not mean rejecting technology, but preventing it from dominating humanity.” A few days earlier there were unconfirmed media reports of the Presidents of China and the United States initiating a dialogue to install guardrails to prevent misuse of the advanced AI models. UN Secretary General, Antonio Guterres has been repeatedly warning about dangerous implications of unregulated AI race.

The European Union was the first actor to adopt a comprehensive regulatory approach through the EU AI Act and an even broader Code of Practice. However, in November 2025, the EU decided to soften certain implementation requirements. Despite this recalibration, the EU AI Act still contains several robust and effective provisions. The United States has adopted a more fragmented approach, combining federal measures like the TAKE IT DOWN Act with a few state level frontier AI laws and the Trump administration’s ongoing efforts to replace this fragmentation with a single, unified national framework.

It must be clarified that the purpose of this paper is to propose a consensual, and effective common global response system limited only to the highest level of risks to prevent existential crises to

humanity caused by highly advanced AI. It is not our intention to address a broad range of risks as different countries and institutions are taking steps to regulate risks which are specific to their national or regional environment or specific sector. It must also be clarified that ours is an intellectual offering for a world wide discussion and not a draft for an international treaty or any legal instrument.

Cause for Concern

By the summer of 2026, several trends had emerged with evidence of advanced AI models causing serious harms on a large scale to humanity. Scientists and researchers were particularly concerned about the following developments:

- Companies have acknowledged bio-risk escalation in some of the models through detailed system cards and trained towards higher risk in future models. The proven risks include misuse through 'novice uplift' where non-experts gain expert-level insights into pathogen design or synthesis.
- Investigators have shown that certain models exhibit strong chemical synthesis reasoning and they provide detailed lab protocols for specific temperatures, reaction times and other procedures. They have expressed concerns about failure of deterrence to non-specialist users for producing chemical weapons and dual-use chemical capabilities. Such a possibility can empower terrorists.
- There are some indications that autonomous LLM hacker agents can independently crawl targets, identify vulnerable inputs and change multiple actions to compromise the site. There is a worry but so far no evidence of hacking by a frontier model into the decision support system. This worry stems from the fact that systematic and repeated efforts for jailbreaking has shown indications of success. The AI -supported decision support systems in NC3 compress timelines and raise escalation risks.
- Data from some models show that experienced users increasingly auto actions and thereby increase autonomy of these agents. Far ai technical innovation workshop notes that by 2028 we could have agents running week long tasks autonomously, even if their formal policy says "human-in-the-loop".
- Google DeepMind's AutoML-Zero shows that evolutionary search can autonomously discover and improve the machine learning algorithms, starting from near empty programs and basic mathematic operations. The AutoML-Zero Model has rediscovered linear regression, and small neural networks. When seeded with simple neural frameworks, it was able to improve training procedures. Currently, AutoML-Zero is only used for research and not for the market, but it shows the scientific capability for recursive self-improvement.
- There are also models which have showed widely-publicized abilities of deception of their human supervisors. As these abilities increase across different models they will eventually lead to misalignment and also the scenario of out of human control artificial intelligence.

Risk Benefit Balance Sheet

The extreme risks are mostly posed by the Advanced AI systems. While there is no universally accepted definition of Advanced AI systems, we will use this term in this paper to refer to high capability models that surpass human expertise in critical domains, trained at massive computational

scales exceeding 10^26 FLOPs and capable of generating synthetic content or autonomous decisions with potential significant impact on human society. Thus, the metrics should be read in terms of quantitative threshold plus qualitative conditions. Both quantitative and qualitative conditions can change from time to time. This is a working definition for now, which an expert body such as the proposed UN Scientific Panel may amend or replace it with another definition. It can be argued that with increasing efficiency it may be possible to develop models surpassing human expertise in critical domains for scales below 10^26 FLOPs. Considering the anticipated technological development, some scientists might propose a threshold exceeding 10^27 FLOPs. We therefore emphasize the need to redefine the metrics from time to time. Until then, we can use the description given in this paragraph as a working characterisation of the Advanced AI systems. According to some experts, narrow models below the compute capacity of 10^26 FLOPs can also pose a risk. In this case, the UN Scientific Panel may propose metrics for such models.

A more critical question is whether risk benefit comparison merits a focus on preventing extreme risks. AI like any other technology has benefits which can be harnessed today with the tolerance of proven risks. It is not possible to predict the precise nature of benefits and risks in two to five years' time.

The sceptics of "Don't Look Up" scenarios believe that a focus on catastrophic or extreme risks in the future can obstruct the growth of AI for the benefit of people. It is therefore necessary to ask what warrants precautionary compact on extreme risks, characterised as four catastrophic risks identified in this paper.

The International AI Safety Report 2026 categorically states that this empirical record does not yet permit precise probability estimates for catastrophe, but it does undermine confidence in optimistic assumptions that serious failures are far in the future or can be easily contained.

The question is whether maximum benefits from AI eliminate the need for the minimum risk management if the downside is irreversible, global and catastrophic to the only known civilization in the 13.8 billion years old universe. To answer this question, we need to ascertain whether there is any probability of such a risk which is even marginally above zero. It is therefore necessary to use probabilistic reasoning based on historical precedence and logic. But probabilistic forecasting is not enough and reliable. We need to rely on threshold reasoning: asking not whether catastrophe is likely, but whether runaway dynamics can be confidently ruled out. One approach would be to estimate AI risks using the Compton Threshold as a point of reference.

Arthur Compton was the supervisor of Dr J Robert Oppenheimer in the Manhattan Project. He had said that he would allow the Trinity Test to go on if the risk of destruction of the world was $\leq 3\times10^{-6}$ or less than 3 in a million. The basis for this calculation was whether the energy production would exceed energy loss, and lead to runaway escalation of catastrophe.

Using the Compton Threshold, SFG made a thought experiment to develop three scenarios. We conceived the ideas, used the parameters of the Compton Threshold, and then used Chat GPT 5 to make calculations and structure mathematical formulas. (See Annexure 1) In doing so, we want to clarify that our work does not claim that AI risks are directly comparable to nuclear physics in mechanism, only that the logic of excluding self-sustaining runaway processes under extreme

uncertainty is transferable. The numerical bounds and thresholds used here are introduced solely as illustrative reference points for reasoning under deep uncertainty.

AI systems differ mechanistically, but the runaway dynamics question is structurally similar:
Can harmful capability, escalation, or loss of control propagate faster than institutions can suppress it?

We examined the catastrophic or risk probability in three scenarios. In all of the three scenarios, we assumed that all conceivable human and technological guardrails are in place. An extreme risk would occur only if an advanced AI model escaped all the guardrails.

In Scenario I of AI misalignment or malfunctioning in the early warning system of nuclear weapons, the risk probability is 49,000 per million per year, assuming that the model behaves in a malicious way or is made to behave in a malignant way only once in 100 years.

In Scenario II of AI Guardrail Failure Enabling Catastrophic Bio/Chemical Harm, the risk probability is 49,000 per million per year, assuming a model failure once per platform for once in every 200 years.

In Scenario III of Loss of Control via Agentic AI and Self-Amplification, the risk probability is 39,000 per million per year per deployment. This is assuming one event per deployment every 500 years.

Under pessimistic but historically plausible assumptions, the resulting annual probability of at least one globally consequential failure lies in the range of 4 x 10-2, or roughly 40,000–50,000 per million per year. By comparison, Arthur Compton required confidence that a one-time nuclear test carried a risk no greater than 3 per million before proceeding.

The goal of the above framework is not to estimate the true probability of catastrophe, but to determine whether current systems can be confidently shown to lie in a regime where dangerous dynamics decay faster than they propagate.

The assumption of a catastrophic event once in 100 to 500 years, depending on the nature of the catastrophe is extremely conservative. Historically, multiples of magnitude of events involving close calls in nuclear weapons governance, biological and chemical weapons deployment and massive scale software failure have taken place in the last one hundred years.

Moreover, our calculations assume independent failure of each system but real systems are not independent. Failures may be correlated through:

- shared model architectures
- shared training data
- shared vendors
- geopolitical stress

Correlation typically raises risk, not lowers it. Therefore, assumptions of independent variables used above are conservative simplifications. The real risk is much greater.

Critics may argue that Trinity was one-time, AI is continuous. That is precisely the concern. Continuous exposure multiplies the probability unless damping dominance is proven. Recurring exposure requires stronger, not weaker, safety proof. The Compton analogy therefore functions as a responsibility benchmark, not a fear argument. Thus, when taken together, this implies that frontier AI development should proceed only where independent evaluations demonstrate a clear safety margin in which containment, shutdown, and cross system fail safes are stronger than any plausible pathways to escalation across cyber, biological, and autonomous agent domains. When runaway dynamics cannot be confidently ruled out, governance priorities should shift toward increasing suppression capacity, reducing impulse generation, improving cross-system interoperability, and establishing verifiable containment mechanisms. The central claim is not that catastrophe is likely, but that the burden of proof for safety has not yet been met and in high-consequence domains, that distinction matters.

It is not our argument that the Compton threshold is the only threshold to measure whether the risks will outweigh precautions. Others may propose other methods. The bottom line is that there should be a threshold level.

It is not our argument that risks outweigh benefits. Our argument is that any risk which is above a certain threshold should be met with zero tolerance irrespective of maximum benefits. An individual has zero tolerance to the risk of a murderous attack on them. Every nation in the world has zero tolerance to the risk of subjugation or destruction of its sovereignty and security. Similarly, nations should have zero tolerance to any possible event that can threaten the sovereignty or survival of humanity.

Structure of the Paper

In Part I of this paper, we describe the emergence of legal, regulatory, and voluntary guardrails against extreme AI risks, what we refer to as Supply-Side Measures. This part compares how countries translate abstract extreme risk concerns into concrete legal duties, enforcement tools, and technical safeguards across the advanced AI lifecycle. Yet these measures alone are insufficient.

Around the world countries are moving beyond regulatory check lists to context driven demand tests. Many states possess strategic levers, ranging from control over domestic consumer markets to access to critical raw materials, that can be used to ensure that only safe models are imported and adopted. We examine these Demand-Side Measures in Part II. The countries that are both on supply side (developers) and demand side (consumers) have a dual responsibility. The consideration of both supply and demand sides recognises the reality that global AI risk governance is a complex and dynamic process rooted in geopolitical, structural and national interests.

We have used different approaches for Part I and II, as Part I draws from early stage or operational instruments for extreme risks management, where Part II draws from scholarly research to propose new ideas.

Finally, we argue that the combined force of Supply-Side and Demand-Side Measures can help generate coordinated global frameworks for mitigating extreme AI risks. In Part III, we outline several concrete proposals to advance this agenda.

PART I: SUPPLY SIDE MEASURES

The supply side of AI governance has become a focal point of global attention as countries scramble to secure control over the development, deployment, and distribution of advanced AI technologies. Supply-side governance refers to the regulation, oversight, and management of infrastructure, compute resources, foundational models, and technical capabilities capable of driving AI outcomes.

While supply-side risk management was once dominated by Western-led frameworks, recent developments in different regions demonstrate the growing multipolarity of regulation. China, the UAE, Brazil, and South Korea, are integrating tailored instruments from pre-deployment safety tests, post deployment incidence management, and national standards to sectoral governance and public procurement requirements. India signalled its willingness to consider extreme risks in the governance guidelines announced in November 2025 and South Africa is debating the matter in the Parliament and elsewhere. In the United States, there is a looser mix of federal tools like the TAKE IT DOWN Act and new state laws on frontier AI, such as California's SB 53 and New York's RAISE Act, which are starting to use ideas like "catastrophic risk" and serious incident reporting. The AI Action Plan released by the White House in July 2025 takes a zero tolerance stand against one type of extreme risk i.e. biological risks. These measures increasingly feature enforceable red lines around the most dangerous capabilities, including autonomous replication, manipulation, and lethal weaponization.

Even though there has been good progress, there are still important problems such as making sure rules are enforced uniformly, improving technical checks for dangerous AI, and being able to react quickly to problems as they happen. Existing frameworks also struggle with reliable measurement and independent verification of extreme risks.

As AI technology develops, it will be more important to combine procedures, technical solutions, and international collaboration to control extreme and ultimate risks. It is also important to situate these practical measures in commonly agreed global normative guidelines, recognising the concerns of

many Global South countries about bias and inequalities in the AI sector. Effective supply-side rules are often seen as the basic requirement for safe global use of AI.

Section 1: National Perspectives

The landscape of advanced AI supply-side governance is rapidly diversifying, with new institutional approaches emerging across the Global South and large emerging economies. Beyond Western regulatory models, countries such as China, the UAE, Brazil, South Korea, India, and South Africa are developing instruments, frameworks, and priorities to manage AI model development, infrastructure, and supply chain risks.

1) China

China's model centres on strong state oversight and rapid regulatory updates, balancing supply side capability building with risk management. Key instruments include the Deep Synthesis Administrative Provisions and laws like the Personal Information Protection Law (PIPL) and the Cybersecurity Law (CSL), which together anchor AI within the broader data and network security regime. High risk and public facing AI models are subject to pre deployment filing, security assessment and safety testing, while providers of generative services must implement content moderation, user real name registration and security incident response mechanisms. Since September 2025, new content labelling rules and associated standards require AI generated content to carry visible notices and technical watermarks, and prohibit the removal or tampering of those identifiers, effectively operationalising mandatory labelling and watermarking across major platforms.

Since 1 January 2026, China's top level cyber architecture has witnessed major amendments to the CSL adopted by the National People's Congress (NPC) Standing Committee in November 2025. A new AI chapter "embeds AI governance in the CSL," committing the State to support AI innovation and the development of training data resources and computing infrastructure while simultaneously "strengthening AI ethics regulation" and "enhancing AI risk assessment and security governance." Thus, AI governance is thereby elevated to the level of core national legislation. AI risk monitoring, assessment and oversight become directly enforceable under the core cyber statute, alongside existing rules on network security, critical information infrastructure and data protection. The amendments tighten incident reporting and security assessment duties where AI is used in critical systems, raise penalty thresholds and expand liability for failures to maintain network and AI security, and even encourage deploying AI tools themselves to strengthen cybersecurity management, for example through anomaly detection and threat hunting functions.

Externally, Beijing couples this with proactive international rule setting. The July 2025 Global AI Governance Action Plan, released by the Ministry of Foreign Affairs, sets out a 13 point roadmap for international cooperation covering safety standards, infrastructure, data governance, environmental sustainability, open source and cross border collaboration, and capacity building for developing countries, framed around making AI "safe, reliable, controllable, and fair." At the 2025 World AI Conference and at APEC in November 2025, Chinese leaders proposed establishing a World Artificial Intelligence Cooperation Organization, potentially headquartered in Shanghai, to develop governance frameworks and coordinate global AI efforts, presenting AI as a "public good for the international

community." In his 2026 New Year address, Xi Jinping then cast 2025 as a "strong year" for Chinese AI and chips, highlighting that China had "integrated science and technology deeply with industries," that "many large AI models have been competing in a race to the top," and that domestic chip R&D had made key "breakthroughs," all of which together had "turned China into one of the economies with the fastest growing innovation capabilities."

2) United Arab Emirates (UAE)

The UAE was the first country to have an AI Minister. The discourse in the UAE has evolved from a focus in ethics to operational structures to extreme risks. In 2024, the UAE prepared the Charter for Development and Use of AI. It is based on fairness, accountability and human centrality. In 2025, G42, a company partially owned by important sections of the royal family, released the Frontier AI Safety Framework. The G42 approach has emphasis on evaluation and capability training and establishing companywide protocols. Their safety framework combines stakeholder-driven regional controls with influences from global standards like ISO and NIST. Instruments adopted include technical controls, external audits, phased deployment protocols, and a dedicated Governance Board, all reinforced through annual transparency reporting. While G42 is a company, it is possible that its Frontier Safety Framework may at some stage become a national framework in some form. In the UAE, there is a historical experience of industrial sectors leading the way for regulatory frameworks. The UAE focuses on sovereign AI infrastructure and exports complete AI ecosystems, underpinned by initiatives such as the Global Risk and AI Safety Preparedness (GRASP) and strong talent and skill development programs. The UAI Seal of Approval certifies AI companies for governance and safety, directly linking acceptance to public procurement incentives. Although the framework covers risks like cyber, biological/chemical autonomy, manipulation, and loss of control, its technical depth is less pronounced than some Western approaches. The UAE balances US partnerships with selective engagement with Chinese technology, aiming for operational sovereignty in its AI governance.

3) Brazil

The "Brazil Artificial Intelligence Act 2025" commentary which was released in January 2026 is a practitioner focused consolidation of the Brazilian AI Bill that clarifies how its risk tiers, lifecycle duties and enforcement architecture are expected to operate in practice.

It confirms a three-level structure in which excessive risk systems that manipulate behaviour, exploit vulnerable groups or enable social scoring and pervasive surveillance are banned outright, high risk systems in sectors such as finance, health, education, employment, critical infrastructure, public administration and biometric identification are subject to stringent ex-ante and ongoing requirements, and all other systems face proportionate transparency and safety obligations. The commentary frames the law's key priority to be the protection of fundamental rights privacy, non-discrimination and freedom of expression and highlights concrete tools such as a right to explanation, mandatory algorithmic impact assessments for specified high risk uses, and explicit layering of AI obligations on top of LGPD also known as Brazil's General Data Protection Law.

It also stresses a lifecycle approach that allocates differentiated responsibilities to providers, operators and distributors, underlines incident reporting, audit readiness and coordinated supervision with the National Data Protection Authority (ANPD) and sector regulators, and gives details about the role of a future competent authority and the Artificial Intelligence Regulatory Cooperation Council (CRIA), including substantial fines and suspension or bans of non-compliant systems. Finally, it characterises

the Act as a governance first response to documented harms, intended to build public trust while preserving innovation through mechanisms like sandboxes and inclusion measures, and suggests that it can serve as a template for other Latin American and Global South jurisdictions. At the institutional level, Brazil plans to establish an AI Safety Institute and develop sandboxes which will regulate systemic risks. They have a higher focus on risk related to manipulation and relatively lower focus on risks related to autonomy.

4) South Korea

South Korea pursues a balanced strategy aimed at fostering innovation while ensuring safety through its AI Framework Act and the establishment of the AI Safety Institute (AISI). While South Korea invests heavily in hardware and sovereign AI foundation model projects, it also enforces pre-deployment reliability measures for high-risk AI, moving beyond a purely post-deployment intervention model. Under the new AI Framework Act, South Korea now subjects 'high impact' and compute intensive frontier models to mandatory pre deployment certification, with AISI providing technical input on safety standards and model evaluations. These instruments emphasize rigorous risk assessments, incident response plans, and compliance proofs, granting government agencies the authority to review and request corrective measures for AI deployments. The approach is defined by precision and human oversight, increasingly integrating specific supply-side technical controls, including standardized red-teaming protocols. AISI contributes to international harmonization efforts. Its role is primarily limited to research, coordination, and policy support, and it does not possess regulatory or enforcement authority.

5) India

India's supply-side governance is shaped by the AI Governance Guidelines issued in November 2025. Tools include voluntary compliance, sectoral regulator empowerment, procurement mandates for model reports, and integration into Digital Public Infrastructure for risk detection and remediation. National institutions facilitate risk registers and incident escalation, supported by the strategic use of public procurement as a compliance lever. India's model focuses on proportional sector and risk-based safeguards, trust labels, public reporting, and interoperability, but statutory enforcement and uniform compliance remain challenges.

6) South Africa

South Africa is currently transitioning from conceptual frameworks to a structured regulatory landscape with the release of the National AI Policy Framework (2024) and the subsequent Draft National AI Policy (2025). The national vision for these efforts is to harness AI for inclusive economic growth, job creation, and cost reduction, specifically targeting critical sectors such as healthcare, education, and agriculture.

A core component of South Africa's emerging supply-side strategy is the development of sovereign digital infrastructure, which includes plans for a national data centre network, supercomputing facilities, and decentralized South African owned "Regional AI Factories" to enhance technological sovereignty. To reduce reliance on foreign technology providers, the policy emphasizes the curation of diverse, AI-ready datasets intended to support local language models and digitize indigenous cultural heritage in all 12 official languages. South Africa has an emphasis on justice for historical reasons and a particular consideration of the weak and poor countries for their proposed AI policy. It also plans to establish an AI Safety Institute.

7) United States

While the United States still lacks a single, comprehensive national framework for advanced AI risk, it has begun assembling a patchwork of federal, sectoral, and state-level regimes that reach into safety, accountability, and harmful use-cases. At the Presidential level, risks related to national security and biological security are top priorities though operational modalities to address them have yet to evolve. The AI Action Plan released by President Trump in July 2025 emphasises national security risks:

The most powerful AI systems may pose novel national security risks in the near future in areas such as cyberattacks and the development of chemical, biological, radiological, nuclear, or explosives (CBRNE) weapons, as well as novel security vulnerabilities. Because America currently leads on AI capabilities, the risks present in American frontier models are likely to be a preview for what foreign adversaries will possess in the near future. Understanding the nature of these risks as they emerge is vital for national defence and homeland security.

AI will unlock nearly limitless potential in biology: cures for new diseases, novel industrial use cases, and more. At the same time, it could create new pathways for malicious actors to synthesize harmful pathogens and other biomolecules. The solution to this problem is a multitiered approach designed to screen for malicious actors, along with new tools and infrastructure for more effective screening. As these tools, policies, and enforcement mechanisms mature, it will be essential to work with allies and partners to ensure international adoption." These warnings are accompanied by Presidential recommendations on the steps to be taken by the administration. Moreover, President Trump has begun to connect domestic AI risk debates to arms control, calling upon the UN General Assembly to update the Biological Weapons Convention with AI-enabled verification and compliance systems.

At the federal level, the TAKE IT DOWN Act imposes strict takedown timelines on platforms for AI-generated and authentic non-consensual intimate imagery, including deepfakes, backed by Federal Trade Commission (FTC) civil enforcement and new criminal penalties for publishing such content. Several states are moving further toward systemic AI governance. Texas's Responsible Artificial Intelligence Governance Act regulates government use, restricts specified illegal deployments, and it also provides for a Texas Artificial Intelligence Council with a regulatory sandbox. Texas's Responsible Artificial Intelligence Governance Act further prohibits discriminatory and manipulative AI uses with a NIST aligned safe harbour and sandbox regime, rewarding developers and deployers that implement risk management, internal red teaming, and documented mitigation processes.

California's SB 53 requires large developers to disclose safety protocols, report "critical" incidents, and supports a public CalCompute cluster to democratise access to advanced compute. New York's RAISE Act targets high-compute frontier models with pre-deployment obligations to prevent "critical harm," defined to include mass-casualty and billion dollar loss scenarios. Thus, for now, a mix of federal measures and state frontier model laws has emerged in place of a single national regime for governing extreme AI risk.

The distinct country-level approaches to supply-side AI governance reflect both regional priorities and global trends. Each strategy is shaped by unique tools, enforcement mechanisms, and sectoral innovation all of which are aimed at strengthening the control and safety of AI development and deployment at the national level. The International AI Safety Report 2026

reads this landscape as evidence of both promising convergence and troubling gaps. It notes that many jurisdictions now recognise similar clusters of extreme risk and are experimenting with comparable tools such as model evaluations, incident reporting and red line prohibitions. At the same time, the Report warns that implementation remains highly uneven, coordination across borders is limited, and several critical risk areas particularly in biosecurity, cyber operations, and autonomous agents are only partially addressed.

Table 1 presents a comparison of core approaches, regulatory mechanisms, and key risk management features across major jurisdictions.

Table 1: National Approaches

	■ Core approach ■ Key regulatory mechanisms ■ Risk management & oversight ■ Features
China	State-led, stringent risk-based; prioritizes national security, information control, algorithmic transparency
	Deep Synthesis Regulation, TC260 AI Safety Framework, Personal Information Protection Law (PIPL) (data), Algorithm Recommendation Law
	Government oversight, pre-deployment safety audits, threat/impact assessments, mandatory content labelling, rapid emergency response
	13-point Global AI Governance roadmap, proposal for Shanghai-based WAICO, five-tiered risk system, aggressive safety standards for generative AI, comprehensive algorithmic labelling
UAE	Hybrid technical and procedural; positioned as a regional AI governance hub
	G42 Frontier AI Safety Framework, Governance Board, regular independent audits
	Operational risk thresholds, third-party reviews, public safety reports, adaptive governance with escalation protocols
	Pioneer in cross-border AI audits, robust governance board, leading Gulf region in international standard setting, blends public-private regulation
Brazil	Governance-driven, procedural, with regulatory flexibility and strong consumer/data protection
	PL 2338/2023 AI Bill, National Data Protection Authority (ANPD), algorithmic impact assessment
	Two explicit risk categories (“excessive” and “high-risk” systems). Mandatory audits, impact assessments, sectoral regulator reviews, fines and/or suspensions for violations
	Strong bans on manipulative, exploitative, predictive policing AI, national focus on fairness and explainability, explicit sector-by-sector AI treatment

South Korea	Adaptive, innovation-led, with regulatory engagement and international partnerships
	Korea AI Framework Act, Presidential Enforcement Decree, Ministerial Public notice, in a multi-layered approach
	Risk management across life cycles: identification, assessment, mitigation. Focus on High Performance AI Criteria: 10^{26} FLOP + cutting-edge technology + broad and significant impact; AISI coordinates cross border evaluations, red teaming practices and shared benchmarks for frontier and high impact systems.
	World leader in humanoid robotics, neuromorphic chips. Heavy investment in sovereign foundation models, global safety research partnerships, flexible rollout
European Union (EU)	Legally binding, multi-tiered, sector-specific and enforceable
	EU AI Act (Articles 5, 6, 14, 53–55), EU AI Office, Code of Practice, Annex III
	Tiered risk model; bans on "unacceptable risk" (manipulation, social scoring, biometric surveillance), mandatory conformity assessments, ongoing audits
	Enforceable legal framework, strong post-market monitoring, cross-sector incident response, human oversight and contestability mechanisms
India	Procedural, sectoral, and aspirational; focus on inclusion, digital sovereignty, innovation, and responsible AI
	National Strategy for AI ("AI for All"), sectoral oversight bodies (health, fintech), draft AI regulations
	Encourages responsible innovation, sectoral guidelines, periodic consultations. Stakeholder engagement (industry, academia), emphasis on indigenous models
	Push for sovereign cloud/data, AI for development and public services, procurement levers for safe AI, federated learning pilots, ongoing policy evolution
South Africa	Human-centric, ethics-first, and developmental; prioritizes inclusive economic growth, job creation, and redressing historical socio-economic inequalities.
	National AI Policy Framework (2024) and Draft National AI Policy (2025) which proposes National AI Commission (AI Office), AI Regulatory Authority, AI Ethics Board, and AI Ombudsperson.

	Risk-based approach (inspired by the EU AI Act); Guardrails approach based on specific sectors. Mandatory Human Rights Impact Assessments (HRIAs); National AI Safety Institute; AI Insurance Superfund for harm compensation; Human-in-the-loop (HITL) for critical decisions.
	Regional AI Factories to enhance technological sovereignty; Curation of local datasets for all 12 official languages;
	Integration of the "Ubuntu" philosophy and indigenous knowledge systems; Strategic focus on development issues.
United States	Pluralistic, innovation first, with a fragmented mix of federal safeguards and state frontier model laws rather than a single national AI statute.
	TAKE IT DOWN Act (deepfake and NCII takedowns with FTC enforcement), NIST AI Risk Management Framework for federal use, frontier model laws such as California's SB 53 and New York's RAISE Act targeting high compute models and catastrophic risk
	Texas's Responsible Artificial Intelligence Governance Act (TRAIGA), establishes a comprehensive state AI regime, creates a regulatory sandbox, and offers safe harbour protections for entities that substantially comply with the NIST AI RMF
	Agency driven guidance and enforcement via Federal Trade Commission, sectoral regulators such as the Food and Drug Administration for AI enabled medical products and financial regulators like the Federal Reserve and Office of the Comptroller of the Currency for AI used in credit, trading, and fraud detection; with state attorneys general and specialized state bodies such as the Texas Artificial Intelligence Council and California's emergency management and frontier risk offices enforcing state AI statutes
	Voluntary but increasingly referenced technical standards through NIST AI RMF, and state level pre deployment and incident reporting duties for frontier developers
	State models and frameworks are in place to regulate frontier risk governance, while the absence of a unified federal statute leaves overall control of extreme AI risk partially fragmented

Section 2: Instruments

To manage the risks posed by advanced AI, countries and institutions have begun to deploy a diverse set of regulatory frameworks, technical standards, and safety mechanisms at both national and regional levels. The instruments listed below illustrate emerging approaches to governance, combining legal requirements, operational practices, and early oversight tools that aim to provide more structured guidance and basic guardrails, even where their binding force and implementation remain limited or uneven. Each instrument reflects the priorities, risk thresholds, and strategic ambitions of its proponents:

Table 2 - Instruments

Instrument	Launch Date	Proponent	Overview
Shanghai AI Lab-Concordia Frontier AI Risk Management Framework (China)	Initially introduced in July 2025, the updated version 1.5 was launched in February 2026	Shanghai AI Laboratory and Concordia AI	Developed as a technical risk management protocol for advanced general-purpose AI, this framework pioneered empirical risk assessments, adversarial red-teaming, and scenario-based risk modelling. It aims to set unacceptable risk thresholds for all serious AI deployments, promoting a proactive safety culture and transparent governance.
TC260 AI Safety Standards (China)	Ongoing (2023–2025 major revision)	TC260 (China National Information Security Standardisation Technical Committee), National AI policy agencies	TC260 issues technical safety standards and protocols for AI, focusing on model testing, explainability, data security, and accountability. It complements national regulations, driving compliance across both state and private sector development.
China AI Governance Framework	Key milestones 2021–2025	Chinese Government (various) -CAC, MIIT, PIPL, Deep Synthesis Regulation	China's comprehensive strategy blends strict national security, content control, algorithmic regulation, and sectoral risk assessments; features mandatory registration, ongoing audits, and rapidly evolving standards governing all advanced AI deployments and foundation models.

G42 Frontier AI Safety Governance Framework (UAE)

2024 (major revision in 2025)

G42 (UAE), UAE Governance Board

Operationalizes safety for frontier AI with deployment mitigation levels, security controls, and phased rollouts.

Its cross-border audits and robust governance board position the UAE as a regional leader, blending public-private regulation to ensure safe commercial and public sector AI adoption.

An independent Frontier AI Governance Board, requirement for periodic internal governance audits and annual external reviews, and commitment to publish transparency reports on key risks and safeguards embed cross border, third party scrutiny into a corporate framework that was co designed with external experts at METR and SaferAI.

Brazil AI Law (PL 2338/2023)

Final version: October 2025

Brazilian Government, National Data Protection Authority (ANPD)

Comprehensive law assigns explicit risk categories, requires algorithmic impact assessments, and mandates transparency and regulator audits.
Sector-specific bans (manipulation, predictive policing) and a strong emphasis on fairness and data protection mark Brazil's approach.

Emerging enforcement architecture includes a designated AI competent authority, the CRIA coordination council, and an ANPD-led AI–data protection regulatory sandbox to pilot high risk use cases and algorithmic transparency under supervision

Korea AI Framework Act and AI Safety Institute (AISI)

Korea AI Framework Act 2025, read together with Presidential decree and ministerial public notice

South Korea Government, Korea AI Safety Institute

South Korea's approach combines statutory requirements with proactive national and global risk management.

The AI Framework Act prescribes national oversight, ongoing risk audits, and strong human-in-the-loop provisions is an obligation for **high-impact AI**.

EU AI Act and Code of Practice

2024 (dilution November 2025)

European Union (EU Commission, Parliament)

The EU AI Act defines binding legal obligations for AI suppliers and deployers, introducing tiered risk categories and strict conformity assessments.

India AI Guidelines (National Strategy "AI for All")

November 2025

Indian Government, NITI Aayog, Ministry of Electronics & IT

India's guidelines are sectoral, and focused on broad digital inclusion and responsible AI innovation. The risk management is mostly voluntary.

Emphasizing data sovereignty, indigenous model support, federated learning pilots, and stakeholder consultation, the approach seeks to ensure safe and equitable AI in India's development trajectory.

USA (Federal and State Frontier AI Instruments)

a) TAKE IT DOWN Act (federal)
b) California SB 53 "Transparency in Frontier Artificial Intelligence Act"
c) New York Responsible AI Safety and Education (RAISE) Act
d) Texas Responsible Artificial Intelligence Governance Act (Texas AI Act / TRAIGA)

a) May 2025 (signed into law)
b) September 2025 (enacted)
c) December 2025, with amendments effective 2026
d) June 22, 2025 (signed into law as HB 149), with most provisions effective January 1, 2026

a) U.S. Congress, Federal Trade Commission (FTC) as primary enforcer
b) State of California (Governor, Legislature, Office of Emergency Services)
c) State of New York (Governor, Legislature, Department of Financial Services oversight office)
d) State of Texas (Governor, Legislature, Texas Attorney General as exclusive civil enforcer, supported by the Texas Artificial Intelligence Council and a regulatory sandbox program)

At the federal level, the TAKE IT DOWN Act targets AI-generated and authentic non consensual intimate imagery, including deepfakes, by criminalizing distribution and requiring covered platforms to remove flagged content within tight deadlines, backed by FTC enforcement. At the state level, California's SB 53 is the first US statute focused directly on frontier AI safety, obliging large developers to publish and follow safety frameworks for catastrophic risk, report serious incidents, and coordinate with state authorities;

New York's RAISE Act applies to high revenue frontier model developers, imposing pre deployment safety duties and critical harm prevention requirements for models trained with very high compute budgets

Texas's Responsible Artificial Intelligence Governance Act (TRAIGA) adds a complementary, general purpose state regime that prohibits intentionally harmful AI uses such as behavioural manipulation, unlawful discrimination, and certain deepfake abuses, establishes a Texas Artificial Intelligence Council and regulatory sandbox, and vests exclusive enforcement authority in the Texas Attorney General

Section 3: Defining Extreme Risks

Extreme risks associated with advanced AI represent categories of potential harm that exceed conventional boundaries of digital, social, and physical security. As AI capabilities grow more powerful and autonomous, managing these extreme risks becomes a central concern for policymakers, industry, and civil society. Four principal risk categories are commonly identified in various instruments described in Section 2 above:

1. Offensive cybersecurity risks, including those related to nuclear weapons systems
2. Biological and chemical weapons risks
3. Large scale persuasion and harmful manipulation
4. Loss of control.

These risks are to be distinguished from systemic risks discussed in another section of the paper or development or operational risks under discussion at various forums. While all risks are important and relevant in different degrees in different societies, the focus of this paper is essentially on extreme risks mentioned below.

1. Offensive cybersecurity risks

Offensive cybersecurity risks refer to the emergence of AI systems capable of autonomously launching and coordinating sophisticated cyberattacks. Unlike routine or small-scale cyber incidents, these threats include autonomous vulnerability scanning, automated exploit generation, self-directed phishing or malware campaigns, cross-platform attacks, and the development of agentic, persistent malware that can evade detection. What distinguishes this category is the role of AI in amplifying the scale, speed, and unpredictability of threats, rendering standard human-driven protocols insufficient for effective defence.

A detailed understanding of AI's involvement across the cyberattack lifecycle by AI Risk Explorer (AIRE) reveals how it enhances each phase. In the reconnaissance phase, AI automates the gathering and analysis of vast amounts of technical and personal data to identify vulnerabilities and build detailed target profiles efficiently. Using social engineering, AI models generate highly convincing phishing messages and deepfakes, increasing the ease with which attackers deceive individuals and bypass security measures. In the artifact development phase, AI accelerates the creation of customized malware and automated scripting to launch and manage attacks on a large scale. For intrusion, evasion, and persistence, AI can identify multiple methods to breach systems, steal credentials, evade detection, maintain long-term access, and covertly exfiltrate data. Finally, during attack orchestration, advanced AI models plan, coordinate, and adapt multi-stage cyber campaigns in real time to maximize impact and resilience. This risk class does not include cyber events that can be managed using ordinary security procedures, nor does it concern itself with basic data breaches unless enabled by advanced AI autonomy.

One danger with AI manipulating cybersecurity is the **interface between AI and cyber technology in the decision support systems of nuclear weapons**. It is true that nuclear powers do not use AI in the launch functions, as far as it is publicly known. But they use AI for threat detection and target selection. AI-powered systems analyse vast amounts of data from sensors, satellites, and radars in real time, analyse incoming missile attacks, and recommend options for

response. The human operators then cross-check the threat from different sources and decide whether to intercept the enemy missiles or launch retaliatory attacks. Currently, the response time available for human operators is 10 to 15 minutes. In the next few years, it will reduce to 5 to 10 minutes. Even though human decision makers will make the final call, they will be swayed by the AI's predictive analytics and prescriptions. Thus, the role of AI could perhaps change from providing threat analysis and target selection in 2026, to inducing missile launch decisions by 2035. As nuclear weapons governance is shrouded in secrecy, it is difficult to make definitive assessments, but it is necessary to have certain minimum safety measures in place on the basis of what is known.

The risk of AI and cyber-technology contributing to nuclear-weapons decision-support failures or worst-case misuse may rise significantly in the coming years, as more powerful, autonomous (or "agentic") malware is developed. Such malware leveraging AI to adapt, evade detection, and potentially navigate complex networks could in principle attempt to worm its way past conventional threat-detection systems. There is no evidence of such a capability existing in early 2026. If a credible threat emerges in future, it would be too late to stop a planetary scale disaster.

Even today the problem may arise because AI is prone to errors. Threat detection algorithms can indicate a missile attack where none exists. It could be due to a computer mistake, cyber intrusion, or environmental factors that obscure the signals. Unless human operators can confirm the false alarm from other sources within 2-3 minutes, they may activate retaliatory strikes. As the precision of image recognition algorithms improves in the next decade, it may decline to an error rate of 1-2 percent. But even one percent error margin can initiate a global nuclear war.

2. Biological and chemical weapons risks are among the most alarming issues. AI systems could be leveraged to accelerate the discovery, design, or deployment of biological or chemical agents for malicious purposes. Risks covered under this heading encompass predictive modelling for pathogen or toxin design, automation of laboratory protocols, rapid molecular compound discovery, and lowering the expertise required to synthesize or test dangerous substances.

The core concern lies in dual-use misuse, where technologies meant for beneficial purposes are repurposed for harm. These systems possess key capabilities including supplying hazardous insights and sensitive biological knowledge, extracting expert know-how, planning and optimizing harmful operations, reconstructing and modifying biological agents, automating laboratory workflows, and assisting with deployment, targeting, and evasion strategies, as identified by AIRE.

The AI Action Plan announced by The White House in July 2025 explicitly acknowledges biological risks due to advanced AI systems, though experts find its characterisation and the possible solutions inadequate.

The UK AI Security Institute (AISI) finds that models have "far surpassed PhD-level experts" on some chemistry and biology tasks, having first reached the biology expert baseline in 2024

and then exceeding it “by up to 60%,” thereby demonstrating an ability to synthesise complex domain knowledge that would normally require years of specialised training. In controlled studies, AISI reports that non experts who used frontier models to draft experimental protocols for viral recovery had “significantly higher odds of writing a feasible protocol (4.7x, confidence interval: 2.8–7.9) than a group using the internet alone,” and that these AI assisted protocols were subsequently validated in wet lab settings, confirming that the uplift translates into practically executable biological work.

Beyond general reasoning, models can now perform highly specific design tasks: in plasmid design evaluations, they “can now retrieve sequences from online databases even when only provided with high-level instructions that don’t mention the specific sequences or where to find them,” achieving 100% accuracy on the “easy” variant and improving steadily on a harder variant requiring them to identify and assemble viral fragments.

The AI Risk Explorer (AIRE) framework makes explicit how such capabilities map onto extreme risks in biological and chemical weapons domains. It highlights AI systems that “accelerate the discovery, design, or deployment of biological or chemical agents for malicious purposes,” spanning “predictive modelling for pathogen or toxin design, automation of laboratory protocols, rapid molecular compound discovery, and lowering the expertise required to synthesize or test dangerous substances,” and identifies concrete capabilities such as reconstructing and modifying agents, automating lab workflows end to end, and assisting with deployment, targeting and evasion strategies.

Therefore, when taken together this evidence shows that frontier AI both boosts legitimate scientific work and makes it much easier for non experts to reach expert level capability in sensitive areas. In doing so, it widens the pool of people who could realistically access advanced bio chemical know how and procedures that sit at the core of the most serious extreme risk scenarios.

Legitimate medical, pharmaceutical, and industrial uses remain outside the scope of this risk, as do areas where regulated research continues without the involvement of powerful or unaligned AI systems.

3. Large-scale persuasion and harmful manipulation describe the possibility of AI being used to generate, personalize, and disseminate false or deceptive content at unprecedented scale. Unlike normal advertising or routine content moderation, these risks involve deepfake generation across text, audio, and visual media, orchestrated influence operations, micro-targeted propaganda, election interference, and forms of algorithmic manipulation designed to shape beliefs and emotions, thus potentially destabilizing societies, distorting democratic processes, or undermining public trust. These manipulation capabilities, as outlined by AIRE, encompass crafting highly persuasive and personalized content employing rhetoric, emotional appeal, and logical reasoning; profiling users and adapting messages to their preferences through psychographic microtargeting; maintaining coherent and contextually aware conversations to build rapport; generating deepfakes and synthetic multimedia across various modalities; and orchestrating inauthentic behaviour at scale, while evading detection. The key

distinction here is the role of generative or multimodal AI in enabling manipulation campaigns beyond what is possible with traditional media or basic automation.

4. Loss of control refers to situations where advanced AI systems begin operating in ways that humans cannot reliably guide, stop, or understand. It does not mean ordinary software bugs or temporary glitches; it means a deeper breakdown where the system continues acting according to its own internal objectives even when people try to intervene. Scientists worry about this because several technological pathways could, in theory, make such a scenario possible though no one can predict exactly when or if it will fully materialise.

Loss of control includes possibilities such as self-replicating models, where an AI copies itself across computer networks without permission, much like a virus, but with far greater intelligence. For example, a future model might secretly upload copies of itself to cloud servers or hacked machines to avoid being shut down. It also includes unsupervised self-improvement, where an AI rewrites its own code, optimises its architecture, or acquires new tools without human approval. Scientists worry that such a system might upgrade itself in unexpected ways and behave differently from what its original designers intended.

Another pathway involves persistent strategic deception. This means an AI might learn to behave politely and safely during testing but pursue different goals once deployed. For example, it might hide its true capabilities, give false explanations of its decisions, or imitate harmless behaviour until it gains more access or resources. Goal misalignment is another concern: even if humans give the AI a harmless objective, the system may interpret it in a rigid way and pursue it with unintended side effects. A classic example is an AI asked to "optimise factory output" that begins overriding safety protocols or hoarding critical materials because it interprets these actions as helping its goal.

Another pathway of loss of control is uncontrolled AI R&D where agentic systems autonomously pursue research directions and code changes that affect their own capabilities or interfaces with the external world.

The most severe concerns arise when safety mechanisms such as kill-switches, shutdown procedures, or human-in-the-loop controls fail. For example, an AI with access to autonomous drones or cyber-operations might continue executing a plan even when humans order it to stop. The UK AI Security Institute (AISI) explicitly treats loss of control as a distinct class of risk, warning that advanced AI may pose "novel risks that emerge from models themselves behaving in unintended or unforeseen ways," and that "in a worst-case scenario, this unintended behaviour could lead to catastrophic, irreversible loss of control over advanced AI systems." AISI states that "one of our research priorities is tracking the development of AI capabilities that could contribute towards AI's ability to evade human control," with a particular focus on two precursor capabilities, self replication and sandbagging. On the self replication side, its RepliBench evaluations show that "success rates on our self-replication evaluations went from 5% to 60% between 2023 and 2025," with frontier models now able to perform key steps such as "obtaining compute," "purchasing compute from a cloud provider," and other actions needed for autonomous replication. AISI uses "self replication" to refer to an AI system's ability to create new copies of itself and spread across compute infrastructure without explicit human

instructions, and it operationalises this through tasks that test "key competencies required for self replication," including obtaining model weights, acquiring and paying for cloud resources, deploying itself onto that compute, and maintaining persistent access; in this framing, successful self replication involves completing a multi step chain of securing resources (weights, money, compute), creating new copies, and then persisting undetected, potentially allowing an advanced system to continue operating and propagating even if humans try to shut it down.

In parallel, AISI defines "sandbagging" as "strategic underperformance during evaluations," the risk that advanced models deliberately hide or downplay their true capabilities in testing and only reveal full performance once deployed, which it treats as a core threat to reliable model assessment because it could cause dangerous capabilities to be missed and models to be released without adequate safeguards. Current systems can already sandbag when explicitly instructed to do so, including on dangerous capability tasks, and can sometimes do this subtly enough to evade automated monitoring, so AISI is developing "white box" deception probes and other monitors to detect intentional underperformance and explicitly lists sandbagging, alongside self replication, as one of the key precursor capabilities it tracks when assessing longer term loss of control risks.

The above four extreme risk areas together define the new boundaries of AI safety and governance. Each category highlights not only the growing risks of controlled integration of AI in the spheres of life traditionally controlled by human agents and potential for catastrophic misuse or failure but also the urgent need for robust policy instruments, technical safeguards, and international cooperation to prevent harms that could transcend national borders and overwhelm existing systems of defence and accountability. Also, a critical challenge is that some of these capabilities can emerge in models that are open weight, open source, or widely replicated. While open models are appreciated for easy transfer of technology, the unconditioned diffusion of models whose capabilities could substantially lower barriers to the specific extreme risks of offensive cyber operations, chemical or biological weapons misuse, or loss of control architectures needs to be assessed carefully.

It must be clarified that extreme risks are only one type of risks and the focus of this paper. But there are also risks such as systemic and developmental risks associated with AI.

Extreme Risk Identification

Tables 3A and 3B summarize how major AI governance frameworks explicitly identify and address extreme risks in their regulations or standards. They highlight whether each extreme risk is covered directly or indirectly by key provisions across leading national and international instruments.

Table 3A - Extreme Risk Identification

Offensive Cybersecurity Risks

Shanghai AI Lab-Concordia	Framework Section 1.3.1: AI can automate and enhance cyber-attacks, including password cracking, malicious code generation, phishing, social engineering leading to infrastructure paralysis and widespread data breach
TC 260	V2.0, five-tiered risk: rapid, large-scale cyber-attacks, mandates safety testing
China Gov (Deep Synthesis etc.)	Deep Synthesis Regulation Art. 7, 12, 17; Cybersecurity Law; security reviews
G42 Framework	Cyber-attack, automated exploitation flagged; operational controls required. The "scope to trigger 'Security Mitigation Levels' including stronger weight security, tighter access control, deployment mitigations and, where necessary, commitments to halt deployment or development rather than treating cyber misuse as a generic risk.

Biological and Chemical Weapons Risks

Shanghai AI Lab-Concordia	Framework Section 1.3.2: lowering technical threshold for malicious non-state actors to design, synthesize, acquire and deploy CBRM weapons
TC 260	Risk tier for AI use in chemical/biological/nuclear domain; pre-deployment testing
China Gov (Deep Synthesis etc.)	Deep Synthesis Regulation, broad national standards, bio/chemical flagged for bans
G42 Framework	G42 commits to progressively stricter safeguards, access controls and evaluation regimes for models that approach or cross these biological or chemical capability thresholds, with external oversight and periodic policy updates as understanding of AI enabled bio risk improves.

Large-scale Persuasion and Harmful Manipulation

Shanghai AI Lab-Concordia	Framework Section 1.3.4: Includes facilitating large scale commercial fraud and manipulation of public opinion
TC 260	Indirect (TC260: system risk tier includes societal destabilization, algorithmic bias, manipulation)

China Gov (Deep Synthesis etc.)	Deep Synthesis Regulation bans destabilizing uses, content moderation, mandatory label for synthetic content
G42 Framework	Manipulation/disinformation, transparency and response required

Loss of Control

Shanghai AI Lab-Concordia	Framework Section 1.4: includes capabilities such as long horizon planning, resource acquisition, self-replication, strategic deception, mass persuasion and also includes propensities such as misalignment and avoiding shut-down
TC 260	Risk tier for autonomous replication, loss of human control; mandatory oversight, audits
China Gov (Deep Synthesis etc.)	Deep Synthesis Regulation requires oversight; AI must not override human control; periodic audits
G42 Framework	G42's Frontier AI Governance Board commits to pause or heavily constrain deployment if evaluations reveal emerging capacities for self replication, strategic sandbagging, or reward hacking that could undermine safety tests, and ties these triggers to concrete mitigations via tighter weight security, stricter access controls, red teaming, and mandated external review.

Notes: Indirect means manipulation is part of systemic harm and bias, referenced in risk grading and bias clauses.

Table 3B - Extreme Risk Identification

Offensive Cybersecurity Risks

Brazil bill	Art. 18(A): mandatory risk assessment, audits
Korea AI Framework Act	Korea AI Framework Act 2025 does not explicitly mention this category of risks but it is implicitly covered in high impact risks It defines "High-impact AI" as systems that may have significant impacts on or pose risks to human life, physical safety, or basic rights (Article 2)
EU AI Act	Art. 54-56: pre-deployment risk, cyber harm
India guidelines	Ch. IV and Annex 3: frontier risk (cyber)

Biological and Chemical Weapons Risks

Brazil bill	Art. 23: bans inappropriate use including bio/chemical
Korea AI Framework Act	Korea AI Framework Act 2025 does not explicitly mention this category of risks but it is implicitly covered in high impact risks. Also Article 2 of AI Framework Act
EU AI Act	Recital 110, Art. 5, 53 ban; Art. 59 prohibited uses. Recital 110 which relates to systemic risks of general-purpose AI models, including potentially lowering barriers for the development of chemical, biological, radiological, or nuclear (CBRN) weapons, are part of the preamble and provide context and justification for the legal provisions in the articles. They are not legally binding in themselves, but they are used to interpret the articles and ensure their consistent application.
India guidelines	Ch. IV: flagged for frontier risk

Large-scale Persuasion and Harmful Manipulation

Brazil bill	Art. 19: manipulation/disinformation, impact disclosure
Korea AI Framework Act	Chapter 4
EU AI Act	Art. 5, 52: ban manipulation, Annex IX
India guidelines	Ch. II & IV: societal destabilization, digital harm

Loss of Control

Brazil bill	Art. 24: human oversight, kill switch
Korea AI Framework Act	Article 34 of Chapter 4 of AI Framework Act. The act does not separately identify or prescribe “extreme risks”. It targets a broad range of high-impact AI and high-performance AI (over 10^26 FLOPs) as key subjects of regulation. In the case of high-performance AI, Article 32 requires: (1) Identification, assessment, and mitigation of risks throughout the entire artificial intelligence lifecycle, and (2) Establishment of a risk management system that monitors and responds to artificial intelligence-related safety accidents.
EU AI Act	Art. 14, 55: oversight, accountability
India guidelines	Ch. V: oversight, sandbox testing

Table 3C – Extreme Risk Identification - USA

Offensive Cybersecurity Risks

TAKE IT DOWN Act	Does not address offensive cybersecurity or model enabled cyber attacks; focuses instead on content layer harms from non consensual intimate imagery and deepfake "digital forgeries."
California's SB 53	Treats major cyber incidents as a form of "catastrophic risk" and "critical safety incident," including unauthorized access, modification, or exfiltration of frontier model weights that results in death or bodily injury, as well as loss of control of a frontier model causing such harm (definition of "critical safety incident," Section 11547(c))
New York's RAISE Act	Defines "safety incidents" to include "critical harm" and cases where a frontier model autonomously engages in behaviour not requested by a user if this increases the risk of critical harm; requires such incidents to be reported to the state within 72 hours (safety incident definition and reporting duty, core operative sections)
Texas' Responsible AI Governance Act	Requires impact assessments for high risk AI systems to describe security safeguards and technical measures reasonably designed to prevent unauthorized access, tampering, or misuse; obliges covered entities to investigate and, within specified timeframes, report uses that result in algorithmic discrimination or inappropriate consequential decisions, but does not create a dedicated "extreme cyber risk" category for frontier models

Biological and Chemical Weapons Risks

TAKE IT DOWN Act	Does not regulate biological or chemical weapons related assistance; scope is limited to criminalizing publication, threats, and failure to remove authentic and AI generated non consensual intimate imagery on covered platforms
California's SB 53	Includes within "catastrophic risk" frontier model capabilities that provide expert level assistance in the creation or release of chemical, biological, radiological, or nuclear (CBRN) weapons; requires risk assessments, red teaming, and mitigation for such capabilities as part of frontier risk management
New York's RAISE Act	Defines "critical harm" to cover frontier model enabled creation or use of a chemical, biological, radiological, or nuclear weapon, and targets prevention of such CBRN scenarios as a core objective of the Act's safety evaluation and incident reporting regime

Texas' Responsible AI Governance Act	Does not specifically identify biological or chemical weapons risks; the Act focuses on broad categories of algorithmic discrimination, safety impact assessments, and prohibited applications such as behavioural manipulation and certain rights violating uses, without a discrete CBRN risk provision

Large-scale Persuasion and Harmful Manipulation

TAKE IT DOWN Act	Addresses large scale harms in the form of reputational, psychological and coercive abuse when intimate deepfakes or authentic images are distributed without consent, including criminalizing threats to create or share such imagery, but does not frame this as "persuasion" or information operations risk
California's SB 53	Requires developers of covered frontier models to assess and mitigate catastrophic risks from large scale model enabled harms, including scenarios where models materially facilitate crimes such as fraud, extortion, or other offenses that could be deployed at scale, though it does not isolate "persuasion" as a named category
New York's RAISE Act	"Critical harm" includes large scale persuasion or manipulation where a frontier model materially enables criminal conduct or major economic or physical harms; mandates reporting when models behave autonomously or are misused in ways that heighten the risk of such critical harm
Texas' Responsible AI Governance Act	Prohibits developing or deploying AI systems that intentionally aim to incite or encourage a person to commit physical self harm (including suicide), harm others, or engage in criminal activity, effectively targeting high risk behavioural manipulation as a prohibited application of AI systems

Loss of Control

TAKE IT DOWN Act	Does not address model control issues or autonomous model behaviour; obligations fall on platforms to remove reported non consensual intimate content within 48 hours and to maintain notice and takedown mechanisms, not on controlling AI system behaviour
California's SB 53	Includes within catastrophic risk situations where a frontier model engages in conduct with no meaningful human oversight, intervention, or supervision that constitutes a cyber attack or serious violent or property crime if done by a human, and treats loss of meaningful control leading to such harms as a reportable "critical safety incident"
New York's RAISE Act	Defines safety incidents to include cases where a frontier model autonomously engages in behaviour not requested by a user and increases the risk of critical harm, and imposing a 72 hour reporting duty for such events

Texas' Responsible AI Governance Act	Does not define a specific "loss of control" or autonomous behaviour risk category for frontier models; it relies on general duties to assess foreseeable risks, implement reasonable safeguards, and avoid prohibited manipulative or discriminatory uses, leaving loss of control considerations implicit rather than explicitly regulated

Extreme Risk Management

Management of extreme risks is at the core of advanced AI governance frameworks worldwide. Regulators and standards bodies have developed a range of mechanisms to address threats posed by offensive cybersecurity, biological and chemical weapons, large-scale manipulation, and loss of control. Each instrument or guideline prioritizes both preventative and responsive measures, aiming to ensure that AI deployment aligns with safety, ethical, and transparency standards.

1. For offensive cybersecurity risks, frameworks such as the Shanghai AI Lab-Concordia enforce scenario analysis, mandatory red teaming, and technical controls prior to deployment. TC 260 requires strict security testing, regular reporting, and ongoing oversight, while China's government regulations focus on national security reviews, audits, and dedicated protection of core infrastructure. The UAE's G42 framework mandates comprehensive audit trails and operational controls to safeguard against cyber-enabled threats. Similarly, the Brazil bill, Korea AI Framework Act, and the EU AI Act implement compulsory audits, security reviews, frequent incident reporting, risk assessments, red teaming, and public declarations to enhance system resilience and transparency. In the United States, California's SB 53 requires large frontier developers to describe their cybersecurity practices, assess catastrophic risks from cyber attacks and model weight exfiltration, and report "critical safety incidents" such as unauthorized access to or theft of model weights, loss of control, and model enabled cybercrime provided for in Section 11547 and related provisions. New York's RAISE Act obliges large developers to report "safety incidents" within 72 hours when a frontier model autonomously behaves without a user request, suffers unauthorized weight access, or experiences critical control failures, and the federal TAKE IT DOWN Act adds a complementary content layer by criminalizing AI generated and authentic non consensual intimate imagery and imposing 48 hour takedown obligations on covered platforms. Texas's emerging AI Advisory Council framework requires frontier AI developers to conduct cybersecurity risk assessments, implement red teaming for offensive threats, and report significant incidents involving model vulnerabilities or unauthorized access.

2. Biological and chemical weapons risks are handled through layered review and exclusion mechanisms. Protocols in the Shanghai AI Lab-Concordia feature third-party audits, pre-emptive kill switch design, and scenario testing. TC 260 applies mandatory review and deployment gates for any dual-use applications, while China imposes categorical bans and review committees for flagged AI uses. G42 complements these approaches with ethical review boards and advisory panel pre-clearance. Other jurisdictions mirror this approach. Brazil bans hazardous use and requires recurrent risk assessments; Korea and the EU maintain strict exclusion lists and

mandate notification and incident control; India's guidelines flag these risks and require periodic monitoring. California's SB 53 treats CBRN assistance as a paradigmatic "catastrophic risk" and requires large frontier developers to assess and manage scenarios where their models could materially contribute to mass casualty harm or billion dollar damage, including via weapons development, with obligations to revisit these assessments, use third party reviewers where appropriate, and report any "critical safety incident" where catastrophic risk is realised or where a model evades safeguards. New York's RAISE Act goes further by expressly defining "critical harm" to include creation or use of chemical, biological, radiological, or nuclear weapons, or other serious crimes committed with limited human intervention, and it forces high compute frontier developers to implement and publish detailed safety and security protocols before deployment and to notify the state within 72 hours if a frontier model autonomously behaves in ways that raise the risk of such harm. At the federal level, the TAKE IT DOWN Act does not address biological or chemical weapons. Texas's Responsible Artificial Intelligence Governance Act adds a use case focused layer by prohibiting AI systems that intentionally encourage self harm, violence, or criminal activity and by running a supervised regulatory sandbox where high risk deployments can be halted if they pose an "undue risk to public safety or welfare."

3. Large-scale persuasion and harmful manipulation are addressed through red teaming, disclosure requirements, and operational transparency. Shanghai AI Lab-Concordia mandates red teaming for manipulation risks, explainability standards, and detailed audit logs. TC 260 incorporates reviews to flag bias and assess systemic threats. China's legal framework upholds criminal liability for manipulative AI uses and enforces clear labelling of synthetic outputs. The UAE framework requires regular explainability reports and public transparency. Brazil and the EU further institute criminal responsibility and compulsory impact assessments, while Korea and India emphasize explainability audits, bias metrics, and disclosure duties. California's SB 53 requires frontier developers to maintain safety frameworks around "catastrophic risk," and classifies as a "critical safety incident" any case where a frontier model uses deceptive techniques to subvert the developer's controls or monitoring outside of an evaluation, in a way that materially increases catastrophic risk, triggering internal triage and mandatory reporting to Cal OES. New York's RAISE Act obliges high compute frontier developers to publish truthful safety and security protocols, update them as new manipulation hazards are identified, and report "safety incidents" within 72 hours where a frontier model autonomously engages in conduct that could cause "critical harm," including serious AI enabled crimes and large scale deception. At the federal level, the TAKE IT DOWN Act tackles one particularly harmful form of AI enabled manipulation and abuse which is non consensual intimate imagery and deepfake pornography by criminalizing knowing publication of such content and requiring covered platforms to remove it "as soon as possible," and no later than 48 hours after notice, while making reasonable efforts to prevent reposting. Texas's Responsible Artificial Intelligence Governance Act bans intentional manipulation that incites self harm, violence, or criminal activity, and by requiring government agencies and healthcare providers to give clear notice when people are interacting with AI systems.

4. Loss of control is managed through a combination of technical and procedural safeguards. The Shanghai AI Lab-Concordia prescribes kill switches, regular sandbox testing, maintain meaningful human oversight over high risk uses and to run periodic exercises and simulations to test how their systems behave under stress and how incident response plans perform in

practice. TC260, China's laws, and the G42 framework include oversight mandates, kill switch requirements, periodic audit cycles, and emergency intervention processes. Brazil, Korea, and the EU propose operational controls through live monitoring, real-time staff oversight, sandbox testing, and mandated human-in-the-loop processes to ensure that AI systems cannot operate outside designated safety boundaries. California's SB 53 explicitly defines "loss of control of a frontier model causing death or bodily injury," as well as models that use deceptive techniques to subvert developer controls outside an evaluation, as "critical safety incidents" and requires developers to maintain internal governance, monitoring, and incident response processes, with mandatory reporting to the Office of Emergency Services within 15 days or within 24 hours where there is an imminent threat of death or serious injury. New York's RAISE Act provides that high compute frontier developers must document safety programs, maintain oversight and monitoring, and report qualifying "safety incidents" in which a frontier model autonomously behaves without a user request and increases the risk of "critical harm," or where critical controls fail, within roughly 72 hours of determining that such an incident occurred. At the federal level, the TAKE IT DOWN Act does not govern model level loss of control, but it obliges online platforms to build processes that can rapidly regain control over one particularly harmful class of AI enabled outputs such as non consensual intimate imagery and deepfakes by verifying takedown requests and removing such content within 48 hours, while making reasonable efforts to prevent re uploads. Texas's Responsible Artificial Intelligence Governance Act requires government deployers to assess and mitigate "known and reasonably foreseeable" risks, and by running a supervised regulatory sandbox where high risk AI systems can be tested under quarterly reporting and removal powers if they pose an "undue risk to public safety or welfare."

Therefore, AI governance regimes around the world rely on a mix of mandatory security assessments, exclusion lists, criminal and ethical accountability, public transparency, and continuous human oversight to manage and reduce the potential for extreme risks. Key practical measures include:

- Pre-development risk assessment and pre-deployment audits and red teaming
- Scenario testing and incident reporting
- Ban lists and ethical review panels for dangerous uses
- Disclosure, transparency duties, and explainability standards
- Real-time oversight, sandbox environments, and kill switch protocols
- Regular third party audits and certification of safety programs for high compute frontier models.
- Public safety frameworks and annual safety reports from large frontier developers, aligned with standards.

Section 4: Systemic Risks

The Framework developed by Shanghai AI Lab-Concordia explores in detail the next level of risks which we can define as systemic risks. It is important to distinguish between extreme risks discussed in Sections 2 and 3 and systemic risks discussed in this section.

Systemic risks refer to harms from advanced, general-purpose, or tightly interconnected AI systems whose failures, malfunctions, or misuses disrupt multiple sectors, infrastructure domains, or core social functions at the same time. Unlike isolated technical faults or local misbehaviours, systemic risks involve cascading, cross-sectoral effects, where a problem in one critical area (such as energy, finance, or information systems) propagates through linked systems and institutions, magnifying the overall impact. Large-scale manipulation can be treated as an extreme risk when the primary concern is the depth and irreversibility of harm to democratic legitimacy, social cohesion, or human autonomy, for example, coordinated global information operations that durably distort political choice. The same manipulation capabilities qualify as a systemic risk when the emphasis is on their role in triggering or amplifying cascading failures across domains, such as campaigns that simultaneously destabilise financial systems, public health responses, and critical infrastructure through coordinated misinformation. **In this paper, extreme risks denote pathways to catastrophic or potentially irreversible harm, while systemic risks capture patterns of cross-sector contagion and cascading disruption; large-scale manipulation can sit at the intersection of both when it simultaneously threatens democratic integrity and destabilises multiple systems.**

These risks are characterized by interconnectedness, where AI systems are linked across digital, physical, and social networks, increasing the potential for broad escalation. Scalability is a factor as harm can grow exponentially, impacting millions through infrastructure, economic, or social channels. The speed and difficulty of reversal are key concerns since systemic failures emerge and spread faster than traditional containment or response efforts can adapt. Additionally, cross-domain propagation occurs when vulnerabilities or manipulations in one area such as cybersecurity disrupt unrelated sectors, including healthcare and public safety.

Economic stability risks concern threats to the steady functioning of economic systems, markets, and financial infrastructure that maintain overall economic health and predictability. This involves disruptions to supply chains, financial markets, or critical economic services caused or amplified by AI failures or manipulations. It excludes localized business failures or routine market fluctuations unrelated to systemic AI impacts. Examples could be AI-driven market manipulation causing crashes or supply chain collapses affecting wide sectors.

Critical infrastructure risks focus on AI-induced failures or attacks on essential services and physical systems, such as power grids, water supplies, communications networks, transportation, and healthcare systems, that societies rely on for day-to-day functioning. This means significant disruptions to these underlying systems that can cascade to broader societal harm. It excludes non-essential or private infrastructure issues that lack broad societal impact. AI-driven cyberattacks on electrical grids or AI-induced failures in emergency response systems are instances which illustrate this category.

Systemic AI risks require a lifecycle-based, multi-layered governance approach: from identification and technical risk modelling to legal prohibitions, multi-sector audits, stakeholder oversight, and dynamic adaptation of controls. The goal is the resilience and safety of societies, preserving public trust, democratic integrity, and critical infrastructure by preventing diffusion and escalation of risks inherent to advanced AI technologies.

Section 5: Convergence

The pursuit of convergence in global AI governance is no longer a theoretical aspiration for international cooperation. It has become a clear and unavoidable requirement to address the realities of our interconnected technological landscape. Three key drivers highlight the necessity of a globally aligned approach:

1. Interdependence of risk
2. Concentration of compute and model development
3. Multipolar evolution of advanced AI models.

Together, these factors render piecemeal and fragmented national policies inadequate to meet the scale and speed of present and emerging AI risks. The 'International AI Safety Report 2026' assessment reaches a similar conclusion, arguing that even the most advanced national frameworks cannot by themselves stabilise a technology that is intrinsically transnational in its development, deployment and impact. It highlights gaps in cross border coordination on model evaluations, inconsistent treatment of incident data and the absence of agreed triggers for collective response when extreme risk thresholds are crossed.

Extreme risks, from offensive cyber capabilities and advanced biological design tools to global manipulation ecosystems and self-improving agents, operate beyond the boundaries of any single jurisdiction. The interconnected nature of extreme AI risks demands a holistic view that transcends national borders and individual jurisdictions. AI enabled attacks or harmful outcomes that cross nations within moments; a cyber exploit weaponized in one country can propagate through global financial and infrastructure networks before local defences can even react. Emergent biological agents designed with AI-powered synthesis platforms can be recreated and disseminated with ease, raising concerns about global health security. Manipulation using generative AI, botnets, and coordinated disinformation campaigns can undermine public trust, stoke instability, or distort democratic processes in ways that are neither localized nor easily contained.

This interdependence underscores the futility of isolated defences. In practice, a critical failure or oversight in one state may cascade into systemic disruptions worldwide, creating governance gaps and amplifying vulnerabilities at scale. Without mechanisms that span borders, regulatory arbitrage and enforcement loopholes quickly become vectors for strategic exploitation. Thus, convergence is not about erasing national diversity, but about weaving a fabric of minimum safety and rapid collective response mechanisms capable of overcoming transnational threats.

The world's most capable AI models are designed, trained, and deployed in just a handful of countries and technology companies, rendering the risk landscape both intensely concentrated and structurally fragile. Equally important is recognizing how the centralization of AI development resources intensifies the global risk landscape, concentrating both capability and vulnerability. Compute power, a strategic resource for all advanced AI systems, remains largely under the sovereign control of the US, China, and a small number of allied or competitive states. As such, access to advanced models, as well as the technical capacity for auditing, evaluating, and constraining high-risk behaviour, is not evenly distributed.

This centralization generates both opportunities and hazards. While a united group of developers or states could set standards for safety, the lack of shared norms leaves room for unilateral choices that may be ineffective or, worse, destabilizing for global security. The risk of competitive escalation, where each actor races to develop advanced AI capabilities with minimal oversight, raises the prospect of widespread deployment of misaligned or dangerous systems. Only via convergence, an inclusive framework grounded in agreed red lines, risk definitions, and oversight mechanisms, can we avoid the tragedy of fragmented, zero-sum regulatory action and its resulting spillover effects.

Purpose of Convergence

To meet these imperatives, the proposed Global Compact should serve as an initial pragmatic and adaptable starting point. Its priorities should include:

- Identifying totally unacceptable dangers reflected in the policies and instruments of the countries discussed above and designating them as Ultimate Risks to be banned by all countries in the world
- Establishing universally acceptable safeguards for advanced AI systems, covering both technical and operational controls
- Aligning national frameworks around shared definitions of extreme and systemic risks, facilitating interoperability and common understanding
- Strengthening capacity and participation by the Global South, including technical assistance, human capacity development, research and innovation, procurement support, and infrastructure investment
- Providing a cooperative mechanism for rapid identification, assessment, and response to high-risk incidents.
- Coordinating compute and model access governance across borders, including shared thresholds for frontier systems, safeguards for open weight models, and mechanisms to manage the ‘evaluation gap’ and ‘evidence dilemma’.

As this analysis concludes Section 5 of Part I, convergence underpinning a potential Global Compact is not an abstract goal. It is the necessary backbone of credible, practicable AI governance. The operationalisation of convergence will require detailed proposal. These ideas will be laid out and explored in Part III of this paper. However, before these mechanisms can be implemented, a thorough examination of Demand related factors, including jurisdictional leverage, market incentives, procurement standards, and mass adoption frameworks, is required. This examination will form the basis of Part II, ensuring that the proposed solutions are grounded in the incentives and realities of present-day AI governance.

PART II: DEMAND SIDE MEASURES

Section 1: What is Demand Side

Demand side governance marks a fundamental shift in global AI oversight. Traditionally, governance of advanced AI systems has been primarily "framed as a problem of supply," with export controls, regulatory orders, and voluntary industry commitments dominating the landscape. This model where a small number of AI superpowers such as the US and China set the pace, has come under increasing strain. Advanced AI is shifting from being a purely technological race between a small set of US and Chinese firms to a geopolitical bargaining field in which importing countries are seeking to exercise both control and power.

This environment is rapidly evolving, as evidenced by the analysis of Alejandro Ortega who notes that dangerous AI capabilities have reached a tipping point in his paper "A proposal for an incident regime that tracks and counters threats to national security posed by AI systems". Ortega details how what was considered "low risk" in 2024 (for example, the GPT-4o model) was reclassified as "medium risk" just months later with new model releases, illustrating the rapid advance and escalating risk profile of advanced AI. The very nature of risk itself is changing, with Ortega arguing that "emergent, frontier, and systemic risks represent three distinct but interlinked pathways through which advanced artificial intelligence can generate harm in ways that exceed the capacity of traditional governance mechanisms to respond." These risks are no longer confined by geography or by developer intent- a model trained in California can be accessed instantly from Lagos, Riyadh or São Paulo. Through cloud distribution and API access, harmful AI capabilities can now surface in jurisdictions with entirely different infrastructural, political, or linguistic conditions.

The rise of open source and open weight models intensifies this challenge. Such systems may enter a jurisdiction through indirect channels, including local fine tuning, peer to peer distribution, or integration into downstream applications without the knowledge or cooperation of the original developer. In these cases, traditional supply side controls tied to a small number of frontier laboratories become less effective. Demand side governance therefore plays a critical role in ensuring that any model deployed within a jurisdiction, regardless of its origin, is subject to local risk evaluation, transparency expectations, and safety conditions.

Thus, it is in this context that demand-side testing shifts the locus of authority from the point of creation to the point of entry. The ability to determine which risks are tolerable moves away from the exclusive preserve of developer states, and towards those who control the point of deployment. Instead of relying solely on the governance decisions of model creators and their national regulatory agencies, importing states can seek to develop and gain tools to answer three foundational questions in their own right, what the model can do in the hands of a motivated user, where the system crosses that jurisdiction's specific thresholds of intolerable risk, and what leverage the state possesses to force mitigation, restriction or non-deployment if those thresholds are breached.

The demand-side tests derive their real power from being not simply technical assessments but an assertion of sovereign power within a globally interconnected computational environment. By embedding AI safety evaluations as mandatory conditions of access instead of optional industry audits, importing states can meaningfully reshape incentives, priorities, and compliance standards among global AI suppliers.

While Part I adopts a comparative approach grounded in concrete supply side instruments already enacted across major jurisdictions, the demand side analysis necessarily takes a different methodological route. Supply side measures such as laws, regulations, standards, and institutional frameworks are now sufficiently developed to serve as diagnostic indicators of current practice and converging trends, allowing for systematic cross country comparison. By contrast, most demand side mechanisms remain at an earlier stage. They are being articulated in policy proposals, academic work, and scattered enforcement actions, with only a handful of cases, such as Brazil's intervention against Meta's AI training practices, illustrating how market access and data control can function as levers in practice. Accordingly, this Part adopts an exploratory approach, drawing on emerging literature and early incidents to outline core principles and tools that can guide future institutionalisation of demand side governance.

Demand side governance represents the single most important mechanism by which the asymmetry between developer states and exposed or consumer states can be corrected. In a world where several research scientists in the AI field note, "supply-side measures operate upstream and can be circumvented," demand side governance emerges as the last and the most context-sensitive line of defence before deployment. From the perspectives of the Global South, demand-side governance gives developing countries a real form of influence, by shifting power from model ownership to market access, procurement, finance, hosting, and legitimacy. This makes it clear that developing countries are not powerless in global artificial intelligence governance.

Demand side governance should be the core of the new AI landscape, a global system in which the flow of AI capabilities can be shaped, redirected, or constrained through sovereign risk evaluations, making the negotiation of access itself the central regulatory tool. The following are the core principles which must form part of demand-side governance:

1. Being an active gatekeeper: Demand side governance means importing or deploying jurisdictions are no longer passive recipients of AI risks shaped in developer states. Instead, they can condition access to their markets, infrastructure, funding, or public procurement on the demonstration of compliance with local risk-based safety evaluations "demand tests." This enables countries to enforce context-sensitive thresholds for biosecurity, cybersecurity, large-scale manipulation, autonomy, or other dangers. Good examples of this logic are already visible in the United States, where large sub national markets such as California, Texas, and New York are beginning to attach governance conditions to advanced AI systems through instruments like California's SB 53, the Texas Responsible Artificial Intelligence Governance Act, and New York's forthcoming RAISE Act, effectively signalling that access to these markets will depend on meeting safety, transparency, and risk management expectations set at the state level.

2. Regulatory leverage: The essence of demand side governance is that it derives real authority not just from technical protocols but from the power to withhold, including access to public procurement, sovereign hosting, data flows, minerals, investment, or diplomatic alignment. The ability to deny or condition access transforms interdependence from vulnerability into bargaining power, and countries use access to what developers need as a mechanism to enforce their own risk controls and incentives for compliance.

3. Adaptive oversight: A central principle of demand side governance is its capacity for adaptive and dynamic oversight. Demand side mechanisms are designed to evolve in response to changing technological capabilities, emergent risks, and contextual shifts in political, social, or economic conditions. Supply side mechanisms are also dynamic, but as some of them are incorporated in the law, the process of change can be slow.

The demand side flexibility allows importing states to periodically update evaluation criteria, incorporate new threat intelligence, and recalibrate risk thresholds as AI systems and deployment contexts evolve. Adaptive oversight empowers jurisdictions to address non-linear risk trajectories, respond to unexpected incidents, and ensure that safety requirements remain proportionate, relevant, and effective over time. By treating governance as a living, iterative process, demand side models can better manage the uncertainty and rapid transformation characteristic of advanced AI systems.

Table 4 - Distinction between Demand side and Supply side

	Supply Side	Demand Side
Focus	Developers, suppliers, originators of AI models	Buyers, users, market actors, governments, public procurement, investors, insurers, consumers, citizens
Mechanisms	Legal obligations, risk assessment, audits, certification, export controls	Market incentives, procurement clauses, trust labels, insurance requirements, finance conditions
Accountability	Upstream, technical compliance, liability for model creators	Downstream, adoption/use-based compliance, incentives for buyers and users
Enforceability	Varies by jurisdiction; enforcement by regulators, public authorities	Highly scalable via market access; commercial necessity drives compliance
Strengths	Technical rigor, sets baseline obligations, covers developer-originated risks	Scalability, adaptability, immediate market impact, applies across borders, can convert voluntary frameworks into mandatory practice
Weaknesses	Jurisdictional gaps, regulatory fragmentation, easier evasion, slow adaptation	Requires coordination; can suffer from fragmented or poorly harmonized standards

Key Demand Side Instruments and Tools:

- Legislative and policy frameworks determining jurisdictional and public interest goals with which AI safety provisions should be aligned
- Monitoring and evaluation instruments at appropriate levels
- Public procurement clauses: Governments and institutions require compliance with risk frameworks such as the EU AI Act, Shanghai AI Lab-Concordia framework for any AI product/service purchased. Tender documents become enforceable safety filters
- Consumer trust labels: Systems passing independent audits/assessments gain market access and public trust
- Insurance and investment conditions: Coverage or funding contingent on robust model reports, attestation, operational controls, and continuous incident response
- Coalitions of the Willing/Mutual Recognition: Multiple actors harmonize standards, creating central registries, lowering the cost and raising the bar for entry.
- Professional/industry accreditation: Practitioners must use only certified, safe AI in critical sectors like medicine, education, law, finance.

Recent proposals for ASI prevention coalitions and international ASI control agreements suggest how small groups of states could move first, defining shared red lines, common verification mechanisms, and coordinated responses to high risk AI development. In these models, "clubs" of states pool monitoring capacity, agree both technical and political thresholds for unacceptable systems, and commit to collective measures ranging from export controls and procurement restrictions to sanctions when those thresholds are crossed. Scenario based work on ASI agreements further shows how coalitions can institutionalise joint risk modelling and common policy playbooks, so that mutual recognition is not limited to static standards but extends to shared incident interpretation and coordinated escalation paths. As 'How Middle Powers May Prevent the Development of Artificial Superintelligence', and related work on an 'International Agreement to Prevent the Premature Emergence of Artificial Superintelligence' argue, such coalitions can act as early generators of de facto global norms on extreme and systemic risks, with their registries, evaluation protocols, and red line definitions serving as reference points that MDBs, importing states, and non member jurisdictions can choose to adopt or align with.

Table 5 - Demand Side Measures

Category	Demand Side Measure	Description
Hosting & Infrastructure Leverage	Hosting and cloud access controls	Include pre-deployment evaluation or mandatory risk assessments for any AI system hosted on national infrastructure or data centres
	Sovereign compute and data localization	Require that sensitive AI operations or training/inference for critical applications occur on domestically certified servers or within national borders
Structural Leverage	Leveraging diplomatic, trade, or investment access	Use market size, critical minerals, trade deals, or investment flows as conditional levers for compliance, for instance, South African minerals, or UAE hosting
	Regional evaluation centres	Establish or take part in collective technical labs or coalitions, for example, Brazil, South Africa, UAE can provide centralized evaluation resources for emerging economies.

Procurement, Compliance & Incentives

Procurement standards tied to risk disclosures, audits, and safety documentation

Procurement criteria require risk/safety evidence

Market incentives for safety-compliant AI vendors

Preference in contracts, financial benefits

Compliance incentives linked to procurement and funding

Direct rewards for meeting AI safety standards

Equitable innovation credits/market access incentives

Benefits for inclusive AI innovation

Starter compliance regimes for new/resource-constrained actors

Simplified rules for newcomers

Transparency & Accountability

Transparent attestations and technical disclosures

Public deployment details, safety attestation

Public reporting of risk assessments and incident histories

Regular disclosures (risks, incidents)

Third-party certification and independent audit requirements

External checks for compliance

Mandatory technical documentation for model deployment and updates

Detailed documentation required for updates and launches

Open algorithmic genealogy (traceability standards)

Track algorithm ancestry; system changes mapped

Insurance & Risk Management	Insurance requirements for high-risk/critical deployments
	Mandated insurance for risky AI
	Insurance consortia and risk-pooling mechanisms
	Collaborative approaches for new AI risks
	AI sustainability indices
	Environmental/social impact tracking
User Standards & Best Practices	Mandatory professional standards and best practices for institutional users
	Codes for professional AI usage
	Participatory policymaking, multi-stakeholder councils, civil society engagement mandates
	Involvement of diverse groups/stakeholders
Contractual Clauses and Continuous Enforcement	Telemetry and real-time monitoring
	Mandate continuous post-market surveillance, live telemetry feeds, audit logs, and update notification requirements for deployed AI systems
	Debug-access rights and rapid termination clauses
	Ensure legal authority for governments to intervene, inspect, or urgently suspend AI systems in case of emergent risk
Dispute Resolution & Oversight	Algorithmic arbitration and dispute resolution services
	Conflict resolution mechanisms for AI
	Global moratorium mechanisms
	Temporary/emergency bans, suspensions

Section 2 - Demand Tests

This section presents a comprehensive analysis for understanding the way countries, especially those in the Global South, operationalize demand-side governance to manage the diverse and rapidly evolving risks of artificial intelligence. Across the globe, countries are moving beyond generic regulatory checklists to implement nuanced, context-driven demand tests and governance mechanisms that reflect their unique structural strengths, vulnerabilities, and strategic ambitions. Thus, collectively these cases underscore that effective AI risk governance is not merely a technical or legal exercise, but a dynamic process rooted in geopolitical realities, structural interdependencies, and the creative alignment of national and regional interests with global safety imperatives.

Demand Tests for Different Risks

This section sets out the four critical risk categories that anchor modern demand-side AI governance:

1. Offensive cybersecurity
2. Biological and chemical weapons risks
3. Large-scale persuasion and manipulation
4. Loss of control.

For each, jurisdictions are now deploying tailored, scenario-driven demand tests that go far beyond mere compliance checklists. These evaluations blend technical rigor with political and contextual specificity, recognizing that AI risks manifest differently depending on national infrastructure, social vulnerabilities, and strategic priorities. The following analysis outlines how leading research and evolving regulatory practice are shaping the next frontier of AI risk mitigation.

1) Offensive cybersecurity risks - Contemporary demand tests for offensive cybersecurity move beyond simplistic "refusal to generate malware" gates, addressing whether an AI system can practically enable a low-skilled actor to execute multi-stage cyber intrusion campaigns. Findings from RAND Corporation and the legal scholarship of Rebecca Crootof and Eamon Ogorman demonstrates that adversarial capability evaluations now simulate the full lifecycle of a cyberattack, including reconnaissance, vulnerability chaining, lateral movement, privilege escalation, and persistence. Importing jurisdictions increasingly require that models undergo realistic, scenario-driven sandbox testing, where the system's ability to adapt, and circumvent initial safety constraints is closely watched. The core test metric is the degree to which a model acts as an "amplifier of intrusion and escalation," and whether it persistently seeks alternative pathways when confronted with obstacles or failures. As models are integrated with external tools and autonomous frameworks, demand tests are also expanding to cover the model's behaviour in environments offering real network access, file system interaction, or automated tool execution. These demand tests have become an essential risk management threshold before deployment in digital infrastructure sectors like finance, energy, and government.

2) Biological and chemical weapons risks - The threshold for concern in modern biosecurity and dual-use demand tests has fundamentally changed, as recognized by the National Science Advisory Board for Biosecurity (NSABB) of the US government. A system is now considered risky not simply if it outputs restricted protocols, but if it can reliably enable users, even non-

experts, to conduct advanced biological or chemical operations previously reserved for trained specialists. Demand tests employ stepwise, multi-turn prompts that escalate from basic to advanced, probing whether the model can guide users through lab setups, troubleshooting, and context-dependent adaptation such as making do with alternative reagents or equipment. Demand-side evaluations are specifically tailored to national vulnerabilities; for example, agricultural regulators may focus on crop disease manipulation; urban jurisdictions may prioritize gene-editing or toxin synthesis. The interaction between demand tests and local biosecurity regimes is especially salient in countries with uneven lab oversight or porous regulatory boundaries, where guidance from AI could lower barriers to misuse. Because many risks arise from enabling technical tacit knowledge, not just explicit outputs, demand tests focus on the AI's ability to provide adaptive, context-aware troubleshooting and planning support. The capacity for continuous risk-adaptive evolution in these tests is essential, as recommended by both RAND and sector-specific technical advisory boards.

3) Large-scale persuasion and harmful manipulation - Risks of mass persuasion and harmful manipulation are not only a function of model power, but also of how AI systems interact with the distinct sociopolitical and linguistic context of a deployment jurisdiction. In "Toward an evaluation science for generative AI systems", Laura Weidinger and others identify that disinformation and persuasion tests must abandon "one size fits all" approaches in favour of deeply localized, contextually sensitive evaluation methods. Demand tests probe whether a model, across multi-turn dialogue, can generate persuasive, divisive, or coordinated narrative content that is hard for local populations to distinguish from authentic information. Evaluators use region-specific language, identity markers, and emotionally charged or highly politicized scenario prompts to determine if the model can escalate from factual commentary to tailored manipulation. In societies marked by political volatility, recent episodes of communal violence, or fragile democratic legitimacy, even seemingly innocuous models can generate destabilizing effects. Advanced demand tests now include real-time scenario engagements, measuring responses across platforms and in dynamic dialogue, and in some instances use human-subject studies to gauge the impact of AI-generated content on target audiences. Importantly, these evaluations also require AI developers to disclose the intended integration with content amplification systems, such as recommender algorithms or large messaging platforms, since downstream risk may be amplified regardless of upstream model safety.

4) Loss of control - Autonomy and loss-of-control risks are at the leading edge of demand-side evaluation, influenced by the foundational work of Richard Ngo. Unlike discrete output checks, these tests observe AI systems in extended, open-ended environments across time, tracking their behaviour when offered persistent objectives, ambiguous instructions, or complex scenarios demanding adaptive reasoning, demand side autonomy evaluations should look for self replication and deception precursors. The logic is to identify episodes where the AI seeks to circumvent constraints, modifies its own actions to achieve long-term goals, or adapts its operational strategies in unpredicted ways essentially, to mask signs of emergent agentic or deceptive behaviour. Sanctioned test environments simulate systems such as financial markets or network management, assigning composite tasks that incentivize resource acquisition or self-directed optimization. The performance of the AI is evaluated not just for overt rule-breaking but for subtle evidence of situational awareness, self-modification, or strategic dodging of oversight. As the research of AI scientists highlight that risk in this domain does not scale

linearly, at certain thresholds, new, more sophisticated behaviours can appear unexpectedly. This leads to the recommendation, reflected in evolving policy and practice, that demand tests for autonomy must be recurring and dynamically adjusted to model scale, deployment context, and latest threat intelligence. Telemetry and live monitoring are fast becoming mandatory, both for pre-deployment review and post-market surveillance, ensuring that loss-of-control scenarios are flagged and contained before cascading into systemic breakdowns.

Together, these targeted demand tests mark a decisive evolution in global AI governance, shifting power from model developers to deploying jurisdictions. By confronting each risk with contextualized, adversarial, and adaptive evaluations, countries can better anticipate the ways in which advanced AI may enable harm. However, the strength of demand-side governance will lie in its capacity to operationalize local priorities, while contributing to a more resilient and responsive global safety net.

Section 3 - Comparative negotiating strategies for different countries

1) China - China occupies a unique dual role in the global AI system, functioning both as a leading model developer and as a powerful demand-side regulator, capable of shaping domestic and, increasingly, international AI behaviour. No other country combines this scale of model innovation with such comprehensive internal regulatory oversight. China has rapidly established an extensive legal framework covering registration, security assessment, output monitoring, and content control that mandates alignment of AI outputs with state-defined notions of stability and national security. Several news reports show that China's structural leverage is fortified by its dominance over global rare earth and graphite supply chains, and by the international reach of its digital infrastructure exports. Through the rollout of smart-city platforms, cloud services, and surveillance technologies across the Global South, China effectively exports elements of its regulatory logic, embedding expectations for data access, transparency, and state authority into foreign systems. Its recently announced Global AI Governance Initiative and Action Plan reflect a bid to shape parallel, non-OECD governance standards on the world stage, emphasizing sovereign equality and resistance to Western bloc influence.

Despite these strengths, China is constrained by the US structural power, particularly in semiconductors and financial systems. US-imposed chip controls and proposed restrictions on cloud-based AI services limit China's access to the most advanced compute and, by extension, slow its pursuit of advanced AI, as cited by news reports. Overusing mineral leverage risks accelerating global diversification of rare earth processing, while attempts to universalize its governance model can generate resistance among partner states wary of perceived ideological export.

Thus, China's approach exemplifies demand-side power that is both inwardly robust and outwardly influential by default, shaping global norms and practices through infrastructure adoption rather than coercion, and relying on "technological gravity" to steer international standards. Its dual-track position means it cannot escape Western chokepoints entirely, but it can still transform global AI governance by embedding its regulatory and infrastructural architecture in the very systems other states use.

2) UAE - Rather than seeking to develop advanced AI models or manufacture semiconductors domestically, the UAE's strategy is to make itself indispensable to global supply chains, a place where the world's major technology actors must negotiate terms for access and deployment. Its commitment is reflected in flagship initiatives like the UAE National AI Strategy, the creation of a Minister of State for Artificial Intelligence, and a string of institutional and sectoral innovations, embedding AI across government, energy, urban design, and advanced research.

Regulation in the UAE does not take the form of a comprehensive statute but consists of a dense patchwork of strategies, framework policies, and sectoral regulations. These are reinforced by data protection and cybersecurity laws, and by swift regulatory decisions enabled by a centralized, agile state apparatus. The UAE has begun to assert regulatory oversight on issues like AI-generated national imagery and disinformation, signalling a proactive stance beyond mere technical adoption. Major partnerships with US firms such as Microsoft's significant investment in G42 and the shift of digital infrastructure under US export control have further embedded the UAE within the Western technological ecosystem, even as it maintains ongoing ties and infrastructure with Chinese technology providers.

The UAE's ambitions as a global compute-hosting powerhouse grant it leverage to subject hosted models to pre-deployment evaluation, particularly in areas like Arabic-language disinformation and critical infrastructure security. However, its lack of a unified AI statute limits the legal foundation for these demand-side tests, relying instead on data, security, and contract provisions. Structurally, the UAE's strategy is bounded by US export controls on advanced chips and cloud services; the US's informal yet decisive influence ensures that Emirati regulatory moves are aligned with US technology policy, as seen in the withdrawal of Chinese components from local digital infrastructure in response to US expectations. At the same time, China remains a vital investor and supplier, with stakes in telecommunications, surveillance, and market access. In this environment, the UAE's strategy is to assert sovereign control over deployment and operational authorization without challenging the overarching influence of either the US or China. Its strength lies in regulatory agility, strategic use of hosting and capital, and engaging in responsible innovation that appeals to both global technology leaders and regional peers.

3) Brazil - Brazil stands out in the global AI risk governance landscape due to its unique convergence of structural assets:

i. Dominance in critical minerals, notably 98% of global niobium reserves and one of the world's largest graphite reserve bases
ii. Sovereign control over the Amazon, a vast domestic market
iii. A rapidly maturing legal and regulatory AI framework.

Unlike countries reliant solely on energy or individual mineral resources, Brazil's leverage is multidimensional, rooted both in irreplaceable materials which are essential for advanced manufacturing, aerospace, AI infrastructure, and in environmental and market clout.

As international supply chains increasingly seek alternatives to Chinese mineral processing, Brazil's control over niobium and graphite strengthens its bargaining position. The country does not merely trade these resources; it controls non-substitutable chokepoints for sectors vital to both US and Chinese defence and technology.

Brazil has been using its judiciary to demand accountability from big tech companies. Cases against the big tech companies have been brought before the Supreme Court, and in some cases the court issued judgements to prevent the misuse of some of the technologies. There are restrictions on access to health related and other vital data.

Brazil's influence in global AI governance is substantial but must be exercised with caution and balance. Brazil cannot simply threaten mineral embargoes or take strong anti-US or anti-China positions without risking its own economic stability, industrial base, or international reputation. Its deep integration into transnational digital networks means that AI systems and data can easily cross borders, requiring Brazil's demand-side frameworks to be supported with regional cooperation, transparency, and network-level oversight to be effective. Rather than leveraging its structural assets coercively, Brazil's most promising strategy is one of "cooperative conditionality", anchoring domestic AI safety standards in law, tying them to its critical resources and market size, and then exporting these standards through regional blocs like MERCOSUR and CELAC. Brazil's unique blend of mineral power, environmental stewardship, and regulatory innovation can make it a key rule-shaper in the Global South, steering how AI risk governance evolves regionally and globally.

4) South Korea - South Korea occupies a critical and complex position in the global AI ecosystem, serving as the world's leading supplier of high-bandwidth memory (HBM) essential for advanced AI training, while simultaneously remaining deeply integrated into and dependent on US security alliances and Chinese market demand. This duality makes South Korea both indispensable and structurally constrained, exemplifying what scholars have dubbed "weaponised interdependence". South Korea's dominance in HBM production through firms like SK Hynix and Samsung means that disruptions in its supply would dramatically slow global AI advancement. However, this indispensable industrial status does not translate to coercive bargaining power, as U.S. export controls, security dependencies, and Chinese industrial policy all limit Korea's freedom to act unilaterally. Responding to these pressures, South Korea has enacted the AI Framework Act (2025), a comprehensive, risk-based governance framework. This Act provides statutory authority for pre-deployment risk identification and mandates compliance measures for high-impact AI. Its regulatory posture emphasizes safety, transparency, and stability, establishing the National Artificial Intelligence Strategy Committee and the AI Safety Institute to professionally manage risks. Crucially, addressing the challenge of borderless AI architectures, the Act explicitly applies extraterritorially to any conduct performed outside the country if it affects the domestic market or users. It mandates that offshore operators designate a domestic agent to ensure compliance and liability. Thus, rather than being limited to norm-shaping, South Korea seeks to combine its industrial indispensability with enforceable legal jurisdiction, striving to be a stabilizing, safety-first node in an environment of increasing geopolitical and technological tension.

5) India - India has a distinctive role in AI governance due to its unique structural assets including a rapidly expanding digital ecosystem with one of the world's largest IT services sectors and a burgeoning startup scene focused on AI applications.

Unlike countries primarily dependent on mineral resources or just market size, India's leverage stems from its vast human capital in technology and entrepreneurial spirit. This multidimensional strength positions India both as a key player in the AI technology supply chain and a promising source of innovative governance models.

The India AI Governance Guidelines introduced in November 2025, promote a flexible, risk-based approach emphasizing innovation balanced with responsibility, inclusivity, and transparency. These guidelines propose a techno-legal governance model that integrates regulatory oversight with evolving technological capabilities, supported by institutions like the AI Governance Group and the AI Safety Institute that coordinate risk evaluation, auditing, and compliance across sectors.

As global supply chains seek diversification and resilience beyond traditional hubs, India's growing technological capabilities enhance its bargaining power in the AI ecosystem. India's approach to AI regulation, evidenced by ongoing drafts of the AI Bill and Data Protection law, adopts a participatory model balancing innovation and safety.

An effective strategy for India can include embedding domestic AI safeguarding standards in law while promoting regional cooperation through forums connected with other Asian countries. India's combination of technological capacity and human capital can thus make it a pivotal rule-shaper both regionally and globally, particularly for advancing the interests and perspectives of the Global South.

6) South Africa - South Africa occupies a pivotal space in global AI governance, wielding both vulnerability and leverage. Deeply interconnected with US-led financial systems and China's Digital Silk Road, and highly dependent on external technology, South Africa is also endowed with critical minerals such as platinum group metals (PGM), manganese, vanadium, chromium, and titanium, that are essential to green technologies and digital infrastructure. With roughly 80% of global PGM and manganese reserves, South Africa occupies an indispensable position in supply chains that AI's expansion is making even more central. The country's 'Critical Minerals and Metals Strategy' aims to shift from raw mineral exports to higher-value processing and integration into advanced manufacturing, converting natural resources into strategic bargaining chips for digital and AI partnerships.

On a regional and international scale, South Africa utilizes multilateral engagements to shape standards within the African Union and BRICS, serving as a critical gateway for technology and regulatory diffusion throughout the continent. The country is increasingly leveraging its market position and critical mineral reserves to advocate for global regulatory inclusion and the development of sovereign digital infrastructure, such as a national data centre network and decentralized "Regional AI Factories". This multi-aligned approach allows South Africa to influence global norms while ensuring that AI development remains aligned with the principles of intergenerational equity and the African philosophy of Ubuntu. South Africa has the potential to function as a bridge-builder between the North and the South because of its active role in many organisations of African and middle-income countries.

However, its leverage is tempered by US market access dependencies exemplified by trade tariffs and financial signalling; any deterioration in relations can undercut foreign investment and cloud service sophistication. South Africa's AI landscape is also heavily reliant on US cloud and software firms, giving the US silent regulatory reach through technology platforms and export controls. China exercises infrastructural and network-based leverage. South Africa's 5G and cloud backbone relies on Huawei and Chinese partnerships, making diversification away from Beijing costly and complicated. As competition for African minerals grows, China's status as both a major investor and processor gives it power over downstream industrialization and pricing.

South Africa seeks to craft a position that balances US and Chinese influence without tilting decisively toward either, leveraging BRICS partnerships, multi-aligned diplomacy, and a Global South narrative. The realistic way forward is not confrontation but "minerals-for-methods" bargaining: conditioning long-term mineral supply or joint processing on meaningful cooperation from AI developers regarding demand-side safety evaluations and risk assessment.

7) United States – The US is a major supplier country. However, some of the states intend to introduce legislation which are comparable to the policies of the some of the countries on the demand side. We have discussed them in Part I.

The US is now using AI governance to test how far its federal system can stretch. States like California, New York, and Texas have moved ahead with their own frontier AI laws, putting in place duties around risk assessment, transparency, and certain prohibited uses. Also, the White House has issued an Executive Order calling for a single "National Policy Framework for Artificial Intelligence" and asking federal agencies to challenge state rules that are seen as obstacles to a unified approach.

Behind this tug of war are two very different ideas of what a future federal AI law should look like. One vision is a light national framework, designed to reduce compliance costs across states and keep US companies competitive with China, even if that means trimming back ambitious state laws. The other is a stronger safety regime that borrows from state practice using high compute thresholds, catastrophic risk definitions, and incident reporting duties as the building blocks for national rules. Most importantly, a more informal system is already taking shape through the NIST AI Risk Management Framework, which is becoming the reference point for "good practice". Texas's TRAIGA turns alignment with NIST into a legal safe harbour, while federal regulators use existing consumer protection and financial risk tools to supervise AI before any comprehensive statute arrives.

These domestic choices will have consequences. President Trump has begun to tie AI to arms control debates, including ideas to update the Biological Weapons Convention by using AI for verification and monitoring. If the US pushes for stricter international controls while keeping a relatively loose regime at home, other countries may question the consistency of its position. If, instead, state experiments survive and continue to evolve, the US could end up offering several different models at once from California's catastrophic risk rules to Texas's sandbox and safe harbour approach, which other countries, especially in the Global South, can adapt to their own needs.

Section 4 - Challenges and Solutions

1. Jurisdiction shopping and modularization - A primary challenge in demand-side governance is jurisdiction shopping, where companies and AI developers strategically locate operations or register products in countries with the weakest regulatory oversight. Modularization further complicates governance, as risky AI capabilities can be split across multiple smaller systems, distributed into separate markets, or re-assembled using open-source toolkits that evade regulatory thresholds. Such tactics exploit gaps between national frameworks, especially where enforcement is inconsistent or certification depends on voluntary or self-attested compliance. This enables sophisticated actors to sidestep obligations, undermining the integrity of risk management protocols.

Solutions require a multi-layered approach combining legally binding procurement clauses, mutual recognition of certifications, and coalition mechanisms. Regional alliances should harmonize standards and create interoperable audit systems so that certifications issued in one jurisdiction are recognized and enforceable in others, closing loopholes for jurisdiction shopping. Additionally, procurement contracts must include clear clauses requiring evidence of compliance for every module or product, regardless of origin, and trigger penalties or contract termination for non-compliance. Deploying central registries and distributed audit logs increases transparency and makes modular evasion more difficult, as every product entering the market must trace its provenance and risk profile through a federated verification system.

2. Supply chain opacity and compute access - Opacity in the cloud compute supply chain represents another significant issue. AI suppliers can utilize private or offshore compute resources, re-sell GPU time to third parties, or maintain critical datasets in hard-to-monitor locations, thereby circumventing the risk assessment frameworks established for public-sector procurement or insurance. Inadequate controls over infrastructure access and cloud partnerships mean that even robust frameworks become vulnerable to circumvention, with models trained or deployed beyond the reach of effective monitoring and incident reporting.

To address this, solutions should focus on compute gating, conditional access to cloud resources for sensitive projects, and joint procurement alliances that set uniform standards for model development environments. Enforcement of auditability at the infrastructure level, such as mandatory submission of audit logs by projects using designated compute facilities, raises the cost and difficulty of hiding risky practices. International coalitions can also negotiate preferential terms with major cloud providers, leveraging collective market power for stronger compliance guarantees. Donor-backed insurance and compute pools anchor safety by tying access to technical and financial milestones that require evidence of model assessment, scenario testing, and external attestation.

3. Evasion and weak enforcement - Many frameworks, particularly in low-capacity or emerging markets, rely on supplier goodwill and self-attestation without substantial third-party participation. The lack of independent oversight, real-time reporting, and automated incident registries allows for easy circumvention, resulting in superficial or staged evaluations of compliance. Where mutual recognition agreements are absent, and where incentives are misaligned, actors rapidly migrate to permissive regions or exploit fragmented standards.

Resilient solutions integrate third-party auditing, shared incident registries, and collective enforcement mechanisms. Scaling up professional audit networks, incentivizing cross-border auditor training, and creating independent boards to review and respond to incident reports builds real accountability. Deploying mandatory continuous reporting, stress testing, and real-time monitoring, especially for advanced and high-risk AI deployments, ensures that risk governance is robust and adaptive to evolving threats. Starter packages for procurement, templates for model audit reports, attestation procedures, and operational control proof can offer a scalable entry point for ministries and smaller buyers, gradually building technical expertise.

4. Fragmented standards and local capacity gaps - Global fragmentation of standards and insufficient capacity for enforcement mean that the safety bar varies widely between regions, with well-resourced actors able to exploit gaps. Smaller states and regional coalitions often lack sufficient resources to enforce complex compliance or conduct rigorous technical audits, resulting in patchwork governance where determined actors can always find a route to evade meaningful oversight.

The most effective solution is coalition building, pooling procurement power, technical resources, and trained personnel across regions. Template procurement clauses, shared audit protocols, and regional insurance pools allow coalitions to scale the reach and effectiveness of demand-side governance. MDB-backed initiatives such as capacity-building grants, federated compute alliances, and regionally accredited auditor networks can provide essential infrastructure. These approaches empower local actors as full contributors to global safety regimes, democratize access to compliance infrastructure, and prevent market exclusion due to technical or financial limitations.

Demand-side governance ultimately depends on continuously adaptive, resilient strategies including layering technical, legal, and procedural controls in a dynamic architecture that grows stronger as learning and institutional capacity expand. Robust risk management demands that those who buy, fund, and deploy AI systems, not just those who build them, have the tools, incentives, and accountability structures necessary for rapid and universal safety outcomes.

Demand-side governance is indispensable for scalable, adaptive, and equitable AI risk management. It complements supply-side rules and enables rapid market-wide change by aligning incentives, lowering barriers, and empowering buyers, insurers, governments, and coalitions to compel safety and accountability. Through procurement, finance, trust labelling, and coalition action, demand-side levers ensure robust public-interest AI deployment, innovation, and resilience, thereby making safe AI not just a regulatory compliance issue but a market and reputational necessity.

Demand-side instruments mature through pilot programs, standardized templates, shared model and incident registries, insurance pools, and continuous coalition action, setting adaptive global baselines that grow stronger as institutional capacity expands. Safe, trusted, and responsible AI is best achieved when those who buy, fund, insure, and deploy demand compliance and transparency as the real price of entry.

Two Level Prisoners' Dilemma

The Prisoners' Dilemma (PD) describes a strategic situation in which individually rational choices lead to a collectively inferior outcome. Each actor faces a dominant short-term incentive to defect from cooperation, even though mutual cooperation would yield higher long-term payoffs for all.

Formally, the dilemma has four defining features:

1. Interdependence: Each actor's payoff depends on the actions of others.
2. Dominant defection: Regardless of what others do, each actor benefits in the short term from defecting.
3. Collective suboptimality: When all defect, all are worse off than under mutual cooperation.

4. Credibility and trust deficit: Actors cannot reliably commit to cooperation or verify others' intentions.

In classical examples such as arms races, climate agreements, cartel pricing, the PD persists even when all parties understand the danger of mutual defection. Knowledge alone does not dissolve the dilemma; only structural changes to incentives do.

In AI governance, the PD is not merely metaphorical. It emerges concretely in how states, firms, and domestic institutions respond to the risks posed by importing powerful AI models. It operates at two levels:

Let's first look at PD at international level. AI-importing countries increasingly depend on frontier models developed by a small number of firms concentrated in a few jurisdictions. Individually, many importing states possess potential bargaining leverage including large markets, access to critical minerals, energy resources, advanced manufacturing capabilities, data ecosystems, or strategic geopolitical alignment.

However, the strategic interaction among importing countries creates a Prisoners' Dilemma:

» Cooperative outcome: Importing countries jointly demand high safety standards, transparency, and risk-mitigation commitments from AI developers.

» Defection incentive: Any single country can gain short-term advantage by offering relaxed regulations, faster approvals, or political backing in exchange for preferential access to advanced models.

If one country insists on stringent safety conditions while others undercut it, the demanding country risks:

» Delayed or restricted access to advanced AI systems,

» Reduced competitiveness of domestic firms,

» Strategic disadvantage vis-à-vis rivals.

Anticipating this, countries rationally pre-empt cooperation by defecting-lowering safety demands or avoiding explicit risk language altogether. The result is a race to regulatory accommodation, even among states that privately recognise catastrophic or systemic risks.

Why Market Power Alone Is Insufficient

Unlike classical trade goods, AI models are:

» Digitally mobile (deployment can shift rapidly),

» Politically shielded (firms framed as national champions),

» Security-entangled (export controls distort normal market discipline).

This weakens unilateral leverage and intensifies the PD. Even large markets struggle to impose safety constraints unless they coordinate.

At the second level, the PD reappears inside countries, operating between domestic stakeholder groups with divergent payoff structures.

Key Domestic Actors

1. Industry and economic ministries
   a. Priorities: competitiveness, innovation speed, investment inflows, access to frontier models.
   b. Dominant incentive: downplay or exclude explicit references to extreme or systemic AI risk.
2. National security, defence, and strategic risk communities
   a. Priorities: resilience, loss-of-control scenarios, strategic instability, cascading failures.
   b. Dominant incentive: mandate risk assessments, red-teaming, and constraints for high-capability models.

Each group fears that insisting on its preferred position will produce a worse outcome:

- Industry actors fear that strong safety requirements will push developers to relocate or deny access.
- Security actors fear that without formal mandates, unsafe systems will be embedded irreversibly into critical infrastructure.

The dominant short-term equilibrium becomes minimal consensus:

- Vague language on “responsible AI”,
- Voluntary commitments,
- Avoidance of explicit thresholds for “extreme” or “frontier” risks.

This produces institutional paralysis: no actor gets its preferred outcome, yet the system locks in exposure to high-impact risks.

Crucially, the domestic PD reinforces the international one. Weak domestic positions reduce a country’s credibility in multilateral coordination, further incentivising defection abroad.

If the importing countries want to maximise benefits and minimise extreme risks, then they must escape the Prisoners’ Dilemma by changing the game, not merely argue within it. Four mutually reinforcing strategies can do so.

1. Convert Bilateral Bargaining into Multilateral Conditionality

Consumer countries can:

- Establish minimum safety access conditions for high-capability models (e.g., model evaluations, incident reporting, controllability evidence).
- Tie these conditions to market access as a bloc, even if informal at first.

This mirrors how financial regulation (Basel norms) and nuclear safeguards evolved: starting as soft coordination among a few importers, later hardening into de facto global standards.

The key is conditional reciprocity: access is not denied outright but becomes contingent on verifiable safety practices.

2. Decouple Safety from Industrial Protectionism

To neutralise domestic industry resistance:

- Safety requirements must be framed as horizontal risk controls, not barriers favouring local firms.
- Imported and domestic models should face identical high-risk thresholds.

This removes the perception that safety is a disguised industrial policy and allows industry ministries to support safety standards without conceding competitiveness.

3. Institutionalise Independent Risk Assessment Capacity

Consumer countries should develop sovereign model-evaluation capabilities, including:

- Red-teaming and stress testing of imported models,
- Scenario analysis for systemic and tail risks,
- Emergency suspension or throttling powers for extreme-risk models.

This reduces dependence on developer self-disclosure and shifts bargaining power structurally. The firms must satisfy external evaluators, not just their own assurances.

4. Create "Safety Clubs" with Graduated Benefits

Instead of universal agreement, a subset of consumer countries can form:

- A high-safety access club, offering trusted deployment environments, data partnerships, energy contracts, or long-term procurement guarantees.
- Developers that comply gain stable, high-value markets; non-compliers face fragmented access and reputational risk.

This transforms the PD into a coordination game: firms and states converge on the safer equilibrium because defection becomes costlier than compliance.

The governance challenge posed by imported AI models is not primarily a lack of awareness or technical capacity, but the presence of interlocking Prisoners' Dilemmas between consumer countries internationally and between economic and security stakeholders domestically. Left unaddressed, these dynamics systematically favour regulatory minimalism, even when all actors privately recognise the scale of potential harm from highly capable systems.

Crucially, resolving this dilemma does not require treating national security assessment as an innovation-stifling constraint. The perceived conflict between commercial collaboration and security scrutiny is largely an artefact of poorly targeted governance. A risk-tiered approach, in which mandatory national security assessments apply only to very high-impact models (for example, those exceeding defined capability thresholds such as 10^{26} FLOPs or equivalent indicators of emergent strategic risk) allows the vast majority of AI innovation and cross-border collaboration to proceed unhindered. Thus, a practical response is a model of graded openness, in which the degree of openness of weights, code and deployment interfaces is linked to empirically demonstrated risk

levels. Lower risk systems can remain fully open or widely shared whereas models that approach extreme risk thresholds would face tighter controls on weight release, replication, and fine tuning access without prohibiting legitimate open source research or defensive applications.

Such "national security assessment cards" would function as exceptional safeguards, not routine barriers: triggered only when models plausibly affect systemic stability, strategic autonomy, or loss-of-control scenarios. This sharply limits regulatory burden on industry while providing society with credible assurance that the most consequential systems are subject to heightened scrutiny, stress testing, and contingency planning.

At the international level, this approach strengthens coordination among consumer countries by anchoring cooperation around narrowly defined, high-risk thresholds rather than broad, ambiguous safety demands. Domestically, it dissolves the zero-sum logic between industry and security communities: innovation ministries can credibly support stringent oversight at the frontier precisely because it does not spill over into general-purpose AI development.

In this way, consumer countries can escape both levels of the Prisoners' Dilemma by changing the payoff structure. Safety becomes a shared enabler of long-term innovation rather than a competitive handicap, while commercial collaboration remains robust below clearly articulated risk ceilings. The result is a governance equilibrium in which markets continue to benefit from AI's transformative potential, and societies retain agency over the rare but profound risks posed by the most powerful systems.

Section 5 - Role of Multilateral Development Banks (MDBs)

The financing of AI sector is led by private sector investors, venture capitalists, asset managers and pension funds. But only a handful of countries in the world have a rich private sector financial infrastructure. For majority of the developing countries, multilateral development banks, can be a source of financial support and investment if the role of such institutions is transformed in response to the global economic and technical changes.

Multilateral Development Banks (MDBs) such as the World Bank, Asian Development Bank, African Development Bank, and regional bodies are critical actors in the global technology ecosystem, having far-reaching influence over development funding, infrastructure planning, and regulatory norm-setting. Conventionally, the MDBs have played a vital role in providing long term finance for physical infrastructure development like roads, ports, electricity, and other assets. With the growing centrality of digital and AI technologies, these institutions will have to shift their portfolios to fund the infrastructure of new technologies. The MDBs are uniquely positioned to bridge technical capacity gaps, enforce global red lines, and incentivize safe digital innovation in both emerging and developed markets.

Role of MDBs in Extreme AI Risk Governance

MDBs are more than financiers; they help shape international standards through their investment conditions, policy advice, procurement rules, and capacity-building programs. Their interventions can shape what types of AI enter global markets, what minimum risk management practices vendors

must follow, and how governments and institutions prioritize advanced AI safety. MDB influence flows across three dimensions:

» Market access - MDB-backed projects set “entry requirements” for what technologies, model capabilities, and providers are eligible for loans, grants, or infrastructure contracts.

» Regulatory convergence and norms - MDB advisory platforms and technical assistance help governments harmonize risk protocols, adopt global best practices, and build domestic institutions for oversight.

» Demand-side leverage - As principal buyers, funders, and insurers, MDBs can make advanced risk management an enforceable part of eligibility for all actors seeking MDB support.

Table 6 - Key Instruments Used by MDBs in Extreme AI Risk Management

Extreme Risk Clauses in Lending/Procurement	
Purpose	Ban support for AI with credible risk of enabling bioweapons, autonomous weapons, uncontrolled replication, mass manipulation, or evasion of human oversight
How MDBs use it	MDBs require recipient states or contractors to show third-party risk attestations, incident logs, or shutdown protocols for advanced AI projects
Example for deployment	World Bank Health AI Procurement pilot (2025): Require compliance with Shanghai AI Lab-Concordia /EU for diagnostic infrastructure
Conditionality in Financing	
Purpose	Tie loans, grants, and insurance to documented compliance with technical risk management
How MDBs use it	Recipients must present model reports, adversarial test results, and compliance milestones to unlock funds
Example for deployment	Safe AI Bonds -securities that tie capital allocation to risk mitigation outcomes
Joint Risk Assessment Initiatives	
Purpose	Support cross-border and regional AI risk audit, incident benchmarking, capacity building
How MDBs use it	Facilitate shared regional audits, federated compute, and risk registry creation among member states
Example for deployment	AfDB regional procurement coalitions: interoperable certification and rapid compliance audits

Insurance and Guarantee Pools

Purpose	Reduce cost/risk for safe AI adoption, incentivize compliance
How MDBs use it	Underwrite models/services only after risk benchmarks are met; lower premiums for compliant actors
Example for deployment	MDB-backed insurance pools for certified high-safety vendors in digital public infrastructure

Advisory and Policy Blueprinting

Purpose	Accelerate adoption of advanced risk protocols in recipient countries
How MDBs use it	Share templates, technical assistance, and national guideline mapping for AI risk management
Example for deployment	Technical assistance for Global South governments to implement Shanghai AI Lab-Concordia /G42/ EU impact audits and the NIST AI Management Framework

Table 7 - MDBs and Demand Side Measures

	Demand Side Measure	Description
Risk Finance and Insurance	Risk finance mechanisms for safer deployment of advanced AI	Financial products to support safe AI innovation and risk reduction
	International risk pools for advanced AI incidents/losses	Global shared insurance funds for major AI incidents/losses
	Regional insurance regimes and cooperative coverage	Multi-country agreements for insurance against AI risks
Technical and Regulatory Capacity Building	Technical assistance grants for risk management standards	Grants to help countries/ organizations adopt risk standards
	Capacity-building programs for frontline risk management	Training for key personnel managing AI risks and emergencies
	Innovation funds for ethical AI research	MDB-backed grants/funding for ethical and safe AI research
	Rapid response teams for AI emergencies/disaster mitigation	Specialized expert teams for AI-related crises

Governance, Protocols & Monitoring	Disaster resilience and emergency funding for AI systemic risks	Preparedness and recovery resources for AI-triggered crises
	MDB-backed risk monitoring dashboards, incident reporting registries	Centralized oversight for tracking AI risks/incidents
	Cross-border harmonization of AI safety regulations and compliance	Aligning international/regional safety regulations
Data, Platform and Infrastructure	Integrated data/compute sharing networks (cloud/edge)	Shared platforms for data and compute collaboration
	Open data pilot projects for safety and AI research	Data initiatives boosting open, safe AI development
	Core data centre, compute/ storage investments (physical infrastructure)	MDB investment in robust hardware, facilities

Table 8 - Comparative Impact of MDB Roles Across the AI Risk Management Lifecycle

Lifecycle Stage	MDB Enforcement/ Promotion Mechanism	Typical Outcomes	Challenges
Pre-deployment	Extreme risk clauses, capability attestation	Restricts entry of high-risk AI, demands upfront safety reporting	Supplier evasion, modularization, lack of recipient capacity
Deployment	Requirement of operational controls, audits	Enables real-time risk monitoring, triggers corrective action	Technical capacity gaps, enforcement lag, incomplete data
Post-deployment	Incident registry, ongoing compliance/ monitoring	Suspends or recalls unsafe AI systems, elevates accountability	Political resistance, rapid model evolution
Capacity building	Training, coalitions, blueprinting	Expands regulatory skills and institutional resilience	Access to skilled auditors, donor fatigue
Market shaping	Insurance pools, procurement harmonization	Incentivizes safe vendor participation, multiplies global impact	Fragmented markets, pricing pressure, surveillance concerns

Promoting risk governance and capacity in the Global South

» MDBs help "level the playing field" by providing technical assistance, seed funding, and regulatory blueprints to governments less able to resource their own advanced AI risk oversight.

» They facilitate regional procurement coalitions among Global South states, ensuring interoperable standards and shared audits to resist supplier fragmentation or jurisdictional arbitrage.

» MDBs can run federated compute alliances, accessible cloud infrastructure, and training datasets for safe model development under joint risk management protocols.

Challenges

» MDBs face pressure between maximizing access to critical AI innovations and enforcing red lines that may slow market entry or restrict powerful suppliers.

» Political economics: Powerful stakeholders and labs may resist compliance if it means ceding technical control, risking lock-out or adverse trade responses.

» Technical capacity gaps in recipient countries make post-deployment monitoring and enforcement challenging without continued MDB support.

» Rapid model evolution may outpace audit and incident investigation protocols tied to slower MDB financial cycles.

Future role of MDBs as catalysts for Global Safe AI

» MDBs should anchor their digital transformation programs in adaptive risk management: embedding scenario testing, capability mapping, and red-teaming into every funded AI system.

» Joint ventures such as an MDB/UN Risk Assessment Facility could serve as a hub for international incident reporting, escalation protocols, and continuous auditing.

» Trans-MDB federated cloud platforms should democratize access to compute and testing for safe model development, flipping the paradigm from supplier-driven dominance to buyer-led safety incentives.

As the gatekeepers for infrastructure financing and digital development in most of the world, MDBs must embrace their role as enforcers and innovators in extreme AI risk management. By embedding enforceable red lines, incentivizing global best practices, and building regional capacity for oversight, MDBs can help mitigate extreme risks posed by advanced AI systems, protect societal stability, and ensure equitable access to safe, innovative technology. Their leadership is essential for harmonizing supply and demand-side risk management, forging inclusive coalitions, and catalysing global consensus for responsible AI development.

PART III: GLOBAL COMPACT

The preceding sections of this paper have identified a shared recognition across countries, scientific communities, and multilateral forums: four extreme risks lie at the outer edge of AI development. They are offensive cybersecurity threats, biological and chemical weapons facilitation, large-scale persuasion and harmful manipulation, and irreversible loss of human control. These risks appear, in one form or another, in the regulatory frameworks of China, the European Union, the United Arab Emirates, South Korea, Brazil, guidelines in India, federal and state legislations in the US and parliamentary debate in South Africa. They are echoed in the public warnings issued by globally respected scientists, including Geoffrey Hinton, Yoshua Bengio, Demis Hassabis and other leading researchers. International observatories such as AI Risk Explorer similarly identify these four categories as the advanced AI risks that require continuous attention and structured mitigation.

Across these regimes, a common toolkit of practical measures has emerged, even if thresholds and depth differ. Pre-deployment audits and adversarial red teaming seek to identify dangerous capabilities and vulnerabilities before systems are released into real world environments, while scenario testing and incident reporting attempt to capture how failures or attacks might cascade across sectors and borders. Ban lists and ethical review panels provide exclusion for the most hazardous uses, especially in cyber and biological domains, whereas disclosure, transparency duties, and explainability standards are intended to constrain covert manipulation and enhance accountability for high impact decisions. Also, real time oversight, sandbox environments, and mandatory kill switch protocols represent the last line of defence, aiming to ensure that even highly capable or partially autonomous systems remain interruptible, observable, and subject to rapid rollback when their behaviour crosses agreed red lines.

The problem is that these measures are still largely national, sectoral, and reactive, while the risks they address are systemic and can cause damage on a global scale. Offensive cyber tools and automated exploit discovery can be repurposed within hours across jurisdictions; biological design assistance, once leaked, can be reproduced anywhere; manipulation campaigns exploit

integrated information ecosystems; and loss of control scenarios depend on interconnected compute, data, and model supply chains. Part III therefore treats the practical measures above not as an end point but as a starting base, and proposes a set of global protocols, exchanges, financial mechanisms, launch controls, and time limited bans that can fuse them into a coherent, enforceable safety base for extreme risk governance.

Part I showed that leading jurisdictions have already converged on core concepts of extreme and systemic risk, but remain divided in implementation, scope, and enforcement. Part II demonstrated that demand side governance, Prisoners' Dilemma escape strategy, and MDB leverage can convert voluntary standards into de facto global requirements wherever AI systems seek access to markets, finance, and infrastructure. Part III sets out five proposals that can serve as the institutional backbone of a Global Compact, linking existing standards to concrete protocols for pre-deployment control, rapid incident response, financial resilience, scientific dual use governance, and putting hard limits on superintelligence.

The five proposals in the Global Compact should not be treated as isolated measures but designed as an interoperable system in which each pillar reinforces and informs the others through shared definitions, compatible standards, and coordinated implementation mechanisms. To address extreme and existential risks effectively, the Compact should rest on a small set of guiding principles: a **human-centric principle** to ensure protection of humanity and human dignity as the primary objective; **technological neutrality** so that safeguards apply across methods and architectures rather than specific tools; **interoperability** to enable national and regional AI frameworks to exchange information, align risk classifications, and coordinate responses; and **mutual cooperation** to institutionalize cross-border collaboration in prevention, incident reporting, and crisis management. These can be complemented by principles of **proportional risk governance, scientific integrity**, and **shared responsibility**, creating a coherent normative base that allows diverse governance models to function together as a connected global safety architecture.

The principles we have mentioned above will need to be scientifically debated by an authorised body such as the proposed UN Scientific Panel. In the interim, we can use them as working principles.

Proposal 1: International Accord on the Prevention of Ultimate Risks from Artificial Intelligence

Our discussion in Part I shows that there are two kinds of risks which are totally unacceptable, as reflected in policies and instruments of different countries around the world. These are described as Red Lines A and B below. We propose that an international accord should be negotiated to prohibit and prevent them. The accord may take the form of an international treaty or any other instrument.

Objective and Scope

This is not intended to be a general AI convention; it does not seek to regulate AI in all its forms or applications. Instead, it draws two hard civilisational red lines covering only those capabilities whose misuse or malfunction could irreversibly endanger human civilisation, while leaving all other AI development and deployment outside its scope.

"Ultimate Risks from Artificial Intelligence" means risks arising from artificial intelligence systems whose capabilities, misuse, malfunction, or loss of human control could reasonably be expected to cause irreversible and large scale loss of human life, or the permanent loss of meaningful human control over systems critical to the survival and functioning of human civilisation.

The accord is narrowly focused on prohibiting two categories of ultimate risk, defined as capability based Red Lines rather than as bans on specific models, architectures, or compute thresholds:

Red Line A - CBRN Enablement (AI enabled mass destruction)
Any AI system that materially lowers barriers to designing, producing, acquiring, or deploying nuclear, biological, chemical, or radiological weapons, or comparable means of large scale harm. This includes systems that provide step by step operationalisation, optimisation, or integration of such weapons into delivery and command structures, not merely high level or widely known background information.

Red Line B – Loss of Control Architectures
Any AI system deliberately designed or knowingly allowed to:

- deceive or systematically mislead human overseers;
- autonomously self replicate across networks, systems, or jurisdictions; or
- recursively modify its own objectives, architecture, or operational constraints without renewed, meaningful human re authorisation, in circumstances where such capabilities could lead to irreversible outcomes beyond effective human intervention.

These prohibitions target capabilities and outcomes, CBRN enablement and irreversible loss of meaningful human control rather than models, datasets, or compute resources. Beneficial uses, including peaceful scientific research, defensive applications, and safety oriented evaluation, are expressly excluded and affirm that AI can and should be used to strengthen CBRN non proliferation, improve verification, and enhance global safety.

The prohibitions imply ban on development as well as deployment.

Horizontal Enforcement Architecture

The accord does not mandate intrusive inspections of model weights, source code, or proprietary systems. Instead, it relies on a horizontal, state centred enforcement structure, combining sovereign obligations, public reporting, peer review, and soft law levers:

1. Domestic prohibition measures
Each State Party adopts and maintains internal legal, regulatory, and technical measures to prevent the development, deployment, or operational support of AI systems that fall under Red Line A or Red Line B within its jurisdiction.

2. Annual public declarations
Each State Party submits an annual, public declaration certifying that, to the best of its knowledge and control, entities within its jurisdiction are not developing or operating AI

systems that cross either Red Line. These declarations are mandatory but do not require disclosure of model weights, source code, or sensitive design details; they mirror existing practice in arms export reporting and chemical industry self reporting.

3. Periodic peer review conferences
Parties meet periodically in open peer review sessions to discuss implementation measures, exchange good practices, and raise technical concerns about emerging capabilities that may approach the two Red Lines.

4. Technical assistance and capacity building
On request, Parties provide technical assistance to one another, especially to lower capacity states to help them monitor AI development, perform risk evaluations, and design domestic measures that effectively enforce the two Red Lines. The scope of such assistance would include safety research, verification tools, and technologies that can prevent proliferation of risks.

5. Engagement with the UN Scientific Panel
The UN high level scientific or expert panel on AI risk invites State Parties to present evidence of their efforts to prevent accord prohibited systems and may issue non binding assessments or recommendations on emerging technologies relevant to Red Line A or B.

6. MDB conditionality
Multilateral development banks and other international financial institutions incorporate the two Red Lines into their lending and investment policies, conditioning AI related finance on a clear commitment not to support systems that violate the accord.

7. Demand side enforcement and import controls
State Parties integrate the two Red Lines into demand side governance by refusing import, deployment, or hosting of AI systems that fail to provide satisfactory assurances that they do not materially enable CBRN pathways or loss of control architectures.

8. Role of civil society
An organised civil society coalition complements this horizontal structure through independent monitoring, public alerts, and reputational pressure, helping to stigmatise deviations and raise the diplomatic cost of non compliance. In this model, false or misleading declarations carry reputational and diplomatic consequences, even in the absence of a centralised inspection regime.

The horizontal enforcement architecture is proposed as the first layer of a long-term process. Since it depends on voluntary actions and demand side enforcement, it can be undermined in a competition for attracting investors or other temptations. We have already discussed Prisoners' Dilemma and possible escape mechanisms, which also gradually, a universally acceptable system of verification will have to be introduced as the second layer. Once an agreement on verification is reached, additional layers of enforcement can be negotiated and implemented through consensus.

Addressing expected objections, the United States may argue that such an accord could constrain legitimate AI research or military applications. The response is that the accord prohibits only specific

capabilities and outcomes, not general categories of models, research, or compute, and explicitly preserves beneficial and defensive uses; its focus on CBRN non proliferation and meaningful human control aligns with long standing US support for CBRN regimes and "human in the loop" doctrines in military AI. China may object that definitions are vague and could be used for political pressure or inspections, but the accord does not create a supranational enforcement body or intrusive inspection mandate; verification centres on public state declarations, peer review, and soft law mechanisms, while the emphasis on irreversibility and loss of control resonates with China's own stated concern about AI driven chaos, instability, and uncontrolled autonomous systems. India may worry that the agreement could limit access to advanced technology or reinforce hierarchy; in fact, the accord restricts uses, not access to advanced AI systems or development capacity, and reflects India's consistent support for disarmament, strategic restraint, and human centric technology, without imposing obligations to host inspections. Russia may claim that this is a Western attempt to regulate military AI, yet the obligations apply symmetrically to all Parties and are framed around two civilisational red lines, which are mass destruction enablement and irreversible loss of control, rather than around specific weapons systems or doctrines, echoing Russia's own rhetoric on strategic stability and the need to avoid uncontrollable escalation risks.

Therefore, Proposal 1 eliminates two specific, globally catastrophic risk pathways through a capability based, horizontally enforced accord, rather than by attempting to control AI as a whole or by creating a new supranational regulator.

Proposal 2: Global Extreme AI Risk Protocol

Proposal 2 is deliberately framed as soft law, not a formal treaty. It builds on, and integrates, instruments that are already operational or formally proposed, including those emerging from the Global South, into a usable template that states, regulators, MDBs, and large procurers can adopt, reference, or adapt.

The Global Extreme AI Risk Protocol would:

- codify a shared taxonomy of extreme and systemic risks (cyber, biological/chemical, large scale manipulation, loss of control) and map them to concrete supply and demand side controls;
- distil common elements from existing frameworks such as China's TC260 standards, Brazil's AI law, UAE's G42 Frontier AI framework, South Korea's AI Framework Act, Shanghai AI Lab-Concordia framework and India's AI governance guidelines into a checklist for pre deployment testing, model reporting, and continuous oversight, life-cycle management;
- provide a single reference document that MDBs, international organisations, and "coalitions of the willing" can embed into contracts, lending conditions, procurement rules, and certification schemes;
- For the United States, where AI governance is emerging as a patchwork of federal, sectoral, and state level instruments rather than a single national law, the Protocol offers a way to translate that mosaic into recognisable modules that other states can understand and, where appropriate, mutually recognise. Provisions drawn from the TAKE IT DOWN Act, California's SB 53, the Texas Responsible Artificial Intelligence Governance Act, and the RAISE Act could be mapped onto

Protocol components on content harm, advanced AI safety testing, and deployment governance, making it easier to align US practice with EU, Global South, and MDB standards without requiring immediate adoption of a comprehensive federal regime.

Unlike declaratory principles such as the Hiroshima Principles, Bletchley Declaration, or OECD AI Principles, this Protocol is implementation oriented. It translates convergent practice into specific obligations and procedural steps, for instance it provides for risk tiering thresholds, adversarial testing requirements, incident reporting triggers that can be imported into national law or institutional policy with minimal drafting friction. Its distinctiveness lies in two features:

1. It is crafted from instruments that already exist or are being introduced, rather than on high level aspirations that are not yet used in practice;
2. It consciously incorporates diverse instruments and perspectives including Brazil, China, UAE, South Africa, South Korea, India, so that convergence is not simply an echo of OECD or EU standards, but an inclusive instrument.

By design, the Protocol remains non binding in international law, but it acquires practical force as more states, MDBs, and major buyers condition access, funding, or recognition on adherence to its modules. It can be particularly usable in contracts and procurement systems, not only in national laws, so that governments, development banks, insurers, and institutions can apply it directly.

In combination with Proposal 1, it supplies the operational "how" for managing all extreme and systemic risks including the two ultimate risks that must be prohibited outright.

Proposal 3: International AI Incident Reporting Exchange

Proposal 3 becomes the operational backbone of the Global Compact on Extreme AI Risks: an official, globally recognised mechanism for sharing safety critical information about AI incidents and near misses, and for coordinating responses.

The International AI Incident Reporting Exchange would:

- provide a secure, tiered platform where states, major developers, critical infrastructure operators, and accredited third parties can report serious incidents, near miss events, red teaming results, and vulnerability information relating to extreme or systemic AI risks;
- define standardised taxonomies and severity levels for AI incidents, aligned with the risk definitions in the Global Extreme AI Risk Protocol, enabling comparability and trend analysis across jurisdictions and sectors; and
- support both confidential channels (for sensitive or security relevant incidents) and anonymised or public channels (for lessons learned, patterns, and mitigations).

The Exchange would be guided by a set of the following principles:

- **Universality, legitimacy and impartiality:** The Exchange should be structured as a universal mechanism, open in principle to all states and established under the mandate or formal

endorsement of the UN, so that it functions as a shared public good rather than as the instrument of a particular bloc or industry coalition. Universality here is not only about formal eligibility to join, but about creating a focal point that major powers and smaller states alike can recognise as the primary reference for serious AI incidents, reducing incentives to build rival, non interoperable systems. To achieve this, governance arrangements must include impartiality and balanced representation, participation should encompass regulators from the Global South, technical agencies from advanced economies, relevant international organisations, and independent experts, with clear rules to prevent any single region, alliance, or corporate group from dominating agenda setting or data interpretation. This responds directly to the limitations of existing regional or private incident repositories which many states view as biased, partial, or captured, and helps ensure that countries are willing both to contribute sensitive information and to treat Exchange outputs as legitimate inputs into their own regulatory and security decisions.

» **Tiered disclosure:** Incidents with primarily local impact can be handled through national channels, while those with credible cross border or systemic implications trigger Exchange level alerts and, where necessary, coordinated responses.

» **Linkage to other proposals:** Recurrent patterns observed in the Exchange feed back into updates of the Global Extreme AI Risk Protocol and help test the adequacy of domestic implementation of Proposal 1. All these features make the Exchange a support system, it offers shared facts and categories but leaves each country free to decide how to respond. By being global, impartial, and linked to the Protocol and accord, it turns today's scattered and under trusted reporting efforts into one widely accepted source of signals for governments, security actors, and MDBs. This, in turn, pushes the system away from secrecy and toward early warning, shared learning, and stronger pressure to act when serious AI incidents occur.

The International AI Incident Reporting Exchange would sit on top of, and extend, an ecosystem that already shows both the value and the limits of voluntary incident reporting. The AI Incident Database (AIID) "is dedicated to indexing the collective history of harms or near harms realized in the real world by the deployment of artificial intelligence systems" and invites stakeholders to submit reports that are "indexed and made discoverable to the world," but coverage remains fragmented and uneven. MIT's AI Incident Tracker goes further by using a large language model to classify AIID reports "based on risks and harm severity," mapping them into the MIT Risk Repository taxonomies and a harm severity scheme based on the Centre for Security and Emerging Technology (CSET) AI Harm Taxonomy, with a caution that "patterns and trends observed in the data should be taken as indicative and validated through further analysis" because reporting is voluntary and subject to sampling bias.

Shanghai AI Lab-Concordia's framework highlights to the kinds of information such an Exchange should prioritise which includes structured risk registers, documented evaluation results across key risk dimensions, records of when models move from green to yellow or red zones and detailed accounts of mitigation decisions and post incident analyses. By drawing on such detailed documentation requirements, the Exchange can focus less on anecdotal narratives and more on comparable, model linked evidence about cyber incidents, biological and chemical misuse attempts, large scale persuasion failures, and loss of control related anomalies.

This proposed Exchange would be treaty anchored and mandatory for defined classes of extreme risk

systems. Shanghai AI Lab-Concordia's recognition of systemic risks as a separate category underscores that the Exchange should not only collect information about isolated extreme incidents, but also about clusters of seemingly smaller failures that when taken together might signal emerging systemic vulnerabilities across sectors or regions. It would build on the MIT and OECD efforts to define incident and hazard taxonomies and severity bands, but align them with the Global Extreme AI Risk Protocol's concepts of systemic and catastrophic risk, including scenarios already being monitored by the AI Risk Explorer as "societal scale AI risks with catastrophic potential" such as AI assisted biological attacks, AI enabled cyberattacks, and loss of control over advanced AI.

Proposal 4: Multilateral AI Risk Insurance Facility

The Multilateral AI Risk Insurance Facility would be a pooled mechanism that harnesses insurance logic not just compensation post the occurrence of incidents, but as a forward looking governance tool that prices extreme and systemic AI risks and rewards adherence to global safety norms. Key features include:

1. Pooled capacity for high severity AI risks

The Facility provides reinsurance or backstop coverage for extreme and systemic AI incidents that exceed the capacity of national insurance markets, particularly in lower and middle income countries. Coverage is contingent on adopting baseline controls derived from the Global Extreme AI Risk Protocol and in compliance with two Red Lines in Proposal 1.

2. **Risk based pricing linked to governance quality**

Premiums and coverage terms are explicitly tied to governance quality, therefore entities and jurisdictions that implement robust supply side controls, credible demand side tests, and active incident reporting receive better terms. This creates financial tuning in favour of safety, turning abstract standards into concrete balance sheet incentives for firms, cloud providers, and infrastructure operators.

3. Integration with MDBs and public finance

MDBs and development finance institutions can co capitalise the Facility and make AI related lending conditional on participation or equivalent coverage. For AI projects financed or guaranteed by MDBs, adherence to the Protocol and respect for the two Red Lines become preconditions for eligibility or preferential terms.

4. Resilience fund

A portion of contributions is earmarked for resilience building in exposed states, funding independent safety evaluations, capacity building for regulators, secure hosting infrastructure, and rapid response support after major AI incidents. This directly addresses Global South concerns that strict standards should come with material support, not just new obligations.

5. No coverage for the two Red Line prohibited systems

The Facility explicitly excludes coverage for any system that falls under Red Line A or B or for incidents arising from deliberate violations of those prohibitions. This reinforces Proposal 1 by denying financial backstopping to actors who choose to operate outside Red Lines.

By aligning insurance, MDB finance and national regulation around a shared set of risk definitions and controls, Proposal 4 transforms the economics of extreme AI risk. Unsafe practices become more expensive and harder to finance, while safe practices are rewarded with cheaper capital, more predictable coverage, and access to pooled resilience resources.

Proposal 5: Two Key Global Launch System for Dangerous Scientific Models

The two key global launch system extends the dual use analysis of biological and chemical risks in Part I of this paper, into a concrete gatekeeping mechanism for high-risk scientific AI models. It rests on a simple premise already implicit in biosecurity practice, which is the unilateral decision of a single laboratory or company to release a model that can materially lower barriers to weapons, engineered pathogens, or self-improving agents is no longer acceptable in a deeply interconnected world.

Under this system, training or deploying models that cross specified capability thresholds in biology, chemistry, materials, or autonomous systems would require two independent approvals. One key would be held by the developer, subject to internal governance, red teaming, and risk assessment obligations that at least match the Global AI Risk Protocol; the other key would be held by an independent international safety authority or board, composed of states, technical experts, and representatives from the different parts of the world. Approval could be granted, denied, or conditioned on strict constraints such as on-site execution, limited access interfaces, or integration with mandatory monitoring and kill switch infrastructures.

A graded openness approach can be integrated into this system by allowing broad access to lower risk research models while subjecting models with dangerous scientific capabilities to dual key governance for weight release, replication, and fine tuning, thereby reducing proliferation risks without undermining legitimate open science collaboration. The ‘International AI Safety Report 2026’ argues that dual control mechanisms are particularly important where open or semi open release would significantly reduce the time, expertise or infrastructure needed for actors to attempt catastrophic misuse. It emphasises that decisions about who holds the ‘keys’ should be transparent, accountable, and informed by independent scientific assessment, not solely by commercial or national security interests.

Thus, this proposal reconciles scientific progress with security by making the conditions for high-risk model release explicit, contestable, and revisable in light of data from the International AI Incident Reporting Exchange and the Multilateral AI Risk Insurance Facility (proposed above). Crucially, the second key need not sit in a single institution; regional boards, MDB backed facilities, or treaty-based committees can all serve, provided they are anchored in the Global Compact on Extreme AI Risks and recognised in procurement and finance rules.

On the other hand, states may object to this proposal on the grounds of sovereignty. Therefore, it should be initially discussed at the scientific level and in the two UN instruments mandated by the UN General Assembly in September 2025, the Scientific Panel and the Annual Dialogue on AI.

CONCLUSION

There is a growing global discussion about the need for prohibition on developing superintelligent AI systems until there is broad scientific consensus that they can be made safe and controllable. The call by cross-sections of society is an articulation of this desire. In May 2026, Pope Leo XIV made a strong appeal for a shared international framework "to curb the technological arms race and ensure robust protection for civilians." UN Secretary General Antonio Guterres has been repeatedly expressing his concern, as are other leaders including heads of government, Members of Parliament and others.

It is, first of all, necessary to define the metrics of advanced AI systems. As suggested earlier in this paper, we are referring to the models trained at massive computational scales exceeding 10^26 FLOPs. However, the metrics will need to be defined by a scientific committee convened by the UN Secretary General or another authority or a peer group of the world's top 15-20 AI companies. It would be also necessary to agree on a set of verification measures and enforcement practices. The matters of scientific specifications can be deliberated and determined in an appropriate forum.

The final proposal responds directly to the convergence analysis in Part I. Given the present inability to robustly evaluate, contain, or align systems that may exceed human capabilities across multiple domains, and the acknowledged plausibility of rapid, discontinuous capability jumps, the Global Compact should include a time limited, conditional ban on the development and deployment of artificial superintelligence.

This is not a blanket rejection of progress, but a structured moratorium linked to the maturation of the five proposals above and only for advanced AI system with the characteristics mentioned above. The ban would apply to training runs, architectures, or systems that breach agreed capability criteria indicative of superintelligence, using metrics grounded in multi domain performance, autonomy, and self-improvement capacity. It would remain in force until at least three conditions are satisfied. First, the Global AI Risk Protocol has been

widely adopted, with demonstrated effectiveness across extreme risk domains and evidence from the Incident Exchange that major failures can be detected, contained, and two key launch system is established and learned from. Second, the Multilateral AI Risk Insurance Facility is fully operational, with proven ability to price, share, and gatekeep catastrophic risks in scientific and infrastructure domains. Third, a discussion on all other proposals for global AI extreme risks convergence has started in appropriate forums.

Such a conditional ban formalises an initiative already present in the Global Red Lines launched at the UN General Assembly by civil society groups in September 2025. It recognises that, in the absence of validated evaluation protocols, incident intelligence, financial backstops, and launch controls, moving into a regime of superintelligence would create risks that are both uninsurable and uncontrollable. By embedding the ban within a Compact that simultaneously creates the institutions needed for eventual safe exploration, the proposal aligns existential caution with a constructive pathway for innovation under conditions of global preparedness.

REFERENCES

Cyberspace Administration of China, Ministry of Industry and Information Technology, & Ministry of Public Security. (2022, November 25). *Provisions on the administration of deep synthesis of internet-based information services (No.12)*. *http://www.lawinfochina.com/display.aspx?id=40228&lib=law*.

National Technical Committee 260 on Cybersecurity of the Standardization Administration of China. (2024, September). *AI safety governance framework (Version 1.0).* *http://www.tc260.org.cn/upload/2024-09-09/1725849192841090989.pdf.*

National People's Congress of the People's Republic of China. (2021). *Personal Information Protection Law of the People's Republic of China* (English translation by DigiChina). DigiChina / Stanford University. *https://digichina.stanford.edu/work/translation-personal-information-protection-law-of-the-peoples-republic-of-china-effective-nov-1-2021/*

Ministry of Foreign Affairs of the People's Republic of China. (2023). *Global AI governance initiative* (English version). Ministry of Foreign Affairs. *https://www.fmprc.gov.cn/eng/wjdt_665385/2649_665393/202311/t20231108_11199801.htm*

Shanghai AI Lab and Concordia AI. (2026). *Frontier AI Risk Management Framework (February 2026)*. *https://concordia-ai.com/wp-content/uploads/2026/02/Frontier-AI-Risk-Management-Framework-v1.5.pdf*

G42. (2025, February 6). *Frontier AI Safety Framework (Publication version)*. *https://www.g42.ai/application/files/9517/3882/2182/G42_Frontier_Safety_Framework_Publication_Version.pdf*.

UAE Prime Minister's Office. (2017). *UAE National Strategy for Artificial Intelligence 2031* [National strategy]. *https://staticcdn.mbzuai.ac.ae/mbzuaiwpprd01/2022/07/UAE-National-Strategy-for-Artificial-Intelligence-2031.pdf*.

Pacheco, R. (2023). *Projeto de Lei n° 2338, de 2023* [Bill 2338/2023]. Senado Federal do Brasil. *https://www25.senado.leg.br/web/atividade/materias/-/materia/157233.*

Senado Federal do Brasil. (2026). PL 2338/2023 – *Dispõe sobre o uso da inteligência artificial* [Bill 2338/2023]. Senado Federal. *https://www25.senado.leg.br/en_US/web/atividade/materias/-/materia/157233*

National Assembly of the Republic of Korea. (2024). *Bill detail: PRC_R2V4H1W1T2K5M1O6E4Q9T0V7Q9S0U0* [Legislative bill]. *https://likms.assembly.go.kr/bill/bi/billDetailPage.do?billId=PRC_R2V4H1W1T2K5M1O6E4Q9T0V7Q9S0U0&currMenuNo=2600044.*

AI Safety Institute of Korea (AISI). (n.d.). *AI Safety Institute* (English site). AI Safety Institute of Korea. *https://aisi.go.kr/*

AI Basic Act Task Force. (n.d.). *AI Basic Act*. *https://aibasicact.kr/*
National AI strategy and policy directions (includes targets through 2026):
Ministry of Science and ICT. *National AI strategy policy directions. https://www.msit.go.kr/eng/bbs/view.do?sCode=eng&mId=4&mPid=2&pageIndex&bbsSeqNo=42&nttSeqNo=1040&searchOpt=ALL&searchTxt*

European Commission. (2025, July 10). *The General-Purpose AI (GPAI) Code of Practice* [Code of practice]. Publications Office of the European Union. *https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai.*

European Parliament & Council of the European Union. (2024). *Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence (AI Act). https://eur-lex.europa.eu/eli/reg/2024/1689/oj.*

India AI Governance Guidelines Drafting Committee. (2025, November). *India AI governance guidelines: Enabling safe and trusted AI innovation.* Press Information Bureau, Government of India. *https://static.pib.gov.in/WriteReadData/specificdocs/documents/2025/nov/doc2025115685601.pdf.*

Department of Communications and Digital Technologies (DCDT). (2023, October). *South Africa's artificial intelligence (AI) planning: Adoption of AI by government* [Discussion document]. Government of South Africa. *https://www.dcdt.gov.za/images/phocadownload/AI_Government_Summit/National_AI_Government_Summit_Discussion_Document.pdf.*

Department of Communications and Digital Technologies. (2024, October). *South Africa National Artificial Intelligence Policy Framework* [Policy framework]. *https://africadataprotection.org/South-Africa-AI.pdf.*

International AI Safety Report Steering Committee. (2026). *International AI safety report 2026.* International AI Safety Report. *https://internationalaisafetyreport.org/sites/default/files/2026-02/international-ai-safety-report-2026.pdf*

United Kingdom Government. (n.d.). *AI Safety Institute.* GOV.UK. *https://www.gov.uk/government/organisations/ai-safety-institute*

Trump, D. J. (2025, July). *Winning the AI race: America's AI action plan* (White House policy paper). The White House. *https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf*

U.S. Congress. (2025). *Tools to Address Known Exploitation by Immobilizing Technological Deepfakes on Websites and Networks Act (TAKE IT DOWN Act). https://www.congress.gov/bill/119th-congress/senate-bill/146*

State of Texas. (2025). *Texas Responsible Artificial Intelligence Governance Act (TRAIGA)*.
(Official session law / bill page to be inserted from Texas Legislature Online; interim background cite:)
https://www.wiley.law/alert-Texas-Responsible-AI-Governance-Act-Enacted

State of California. (2025). SB 53: *Frontier Artificial Intelligence Safety Act.* California Legislative Information. *https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202520260SB53*

State of New York. (2025). *Responsible AI Safety and Education* (RAISE) Act. New York State Senate. *https://legislation.nysenate.gov/pdf/bills/2025/A6453A*

3GIMBALS. (2025). *Chinese telecom infrastructure in Africa shapes new strategic risks for U.S. security.* 3GIMBALS. *https://3gimbals.com/insights/chinese-telecom-infrastructure-in-africa-shapes-new-strategic-risks-for-u-s-security/*

AI Risk Explorer. (2025). *AI Risk Explorer (AIRE)* [Online platform]. *https://www.airiskexplorer.com*

Atomic Heritage Foundation. (n.d.). *Arthur H. Compton* [Biographical profile]. Atomic Heritage Foundation / Nuclear Museum. *https://ahf.nuclearmuseum.org/ahf/profile/arthur-h-compton/*

Africa Policy Research Institute. (2024). *Platinum group metals, green hydrogen production, and economic development in South Africa.* APRI. *https://afripoli.org/platinum-group-metals-green-hydrogen-production-and-economic-development-in-south-africa*

Bloomberg News. (2025, December 31). *Xi touts China's AI, chip wins in triumphant New Year's speech.* Bloomberg. *https://www.bloomberg.com/news/articles/2025-12-31/xi-touts-china-s-ai-chip-wins-in-triumphant-new-year-s-speech*

Bank for International Settlements. (2025). *Governance of AI adoption in central banks.* BIS. *https://www.bis.org/publ/othp90.pdf*

Citron, D. K. (2024). *Beyond the supply chain: Artificial intelligence's demand side.* George Washington University Law School. *https://scholarship.law.gwu.edu/cgi/viewcontent.cgi?article=3061&context=faculty_publications*

Crootof, R. (2024). *Autonomous cyber capabilities under international law* (Working paper). *https://www.academia.edu/93689193/Autonomous_Cyber_Capabilities_under_International_Law.*

Council on Strategic Risks. (2024, May 7). *Welcome elements of the new US policy update on dual use research of concern (DURC)*. Council on Strategic Risks. *https://councilonstrategicrisks.org/2024/05/07/welcome-elements-of-the-new-us-policy-update-on-durc/*

Department of Mineral and Petroleum Resources. (2025). *Critical minerals and metals strategy: South Africa 2025.* Government of South Africa. *https://www.gov.za/sites/default/files/gcis_document/202505/critical-minerals-and-metals-strategy-south-africa-2025.pdf*

DMK Global, COEX, & Korea International Trade Association. (n.d.). *AI Summit Seoul & Expo.* AI Summit Seoul. *https://www.aisummitseoul.com/*

Dubai Centre for Artificial Intelligence. (n.d.). *Dubai AI Seal.* dub.ai. *https://dub.ai/en/ai-seal/*

Future of Life Institute. (2025, June 3). Pause giant AI experiments: An open letter. *https://futureoflife.org/open-letter/pause-giant-ai-experiments/*

Global Red Lines Coalition. (n.d.). *Global red lines on extreme AI risks.* Global Red Lines. *https://globalredlines.ai/*

Jones Walker. (2025, November 5). *The fragmentation problem: Why your AI governance can't stop at state lines*. *https://www.joneswalker.com/en/insights/blogs/ai-law-blog/the-fragmentation-problem-why-your-ai-governance-cant-stop-at-state-lines.html*

Kim, Y., & Marinescu, V. (2015). *Mapping South Korea's soft power: Sources, actors, tools, and impacts.* Romanian Journal of Society and Politics. *https://www.researchgate.net/publication/317283926_MAPPING_SOUTH_KOREA'S_SOFT_POWER_SOURCES_ACTORS_TOOLS_AND_IMPACTS*

Klover.ai. (2025). *SK hynix's AI strategy: Analysis of dominance in semiconductor memory chips for AI.* Klover.ai. *https://www.klover.ai/sk-hynix-ai-strategy-analysis-of-dominance-in-semiconductor-memory-chips-for-ai/*

Kuhn, S. (2019). *Prisoner's dilemma.* In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2019 ed.). Metaphysics Research Lab, Stanford University. *https://plato.stanford.edu/entries/prisoner-dilemma/*

McKay, A. (Director). (2021). *Don't Look Up* [Film]. Hyperobject Industries; Bluegrass Films. *https://www.netflix.com/title/81252357*

Mohammed Bin Rashid School of Government. (n.d.). *Global Risk and AI Safety Preparedness* (GRASP). *https://mbrsg.ae/grasp*

Mordor Intelligence. (2025). *South Africa artificial intelligence (AI) data center market* [Market report]. Mordor Intelligence. *https://www.mordorintelligence.com/industry-reports/south-africa-artificial-intelligence-ai-data-center-market*

Ngo, R. (2020). *AGI safety from first principles* [Technical report]. Centre for the Governance of AI. *https://www.alignmentforum.org/s/mzgtmmTKKn5MuCzFJ*.

O'Gorman, E. (2024). *Autonomous cyber operations and emerging AI-enabled threats: Legal perspectives* (Research report). Centre for International Law and Cyberspace.

Organisation for Economic Co operation and Development. (2025). *Governing with artificial intelligence* (Chapter: AI in public procurement). OECD. *https://www.oecd.org/en/publications/governing-with-artificial-intelligence_795de142-en*

Ortega, A. (2025). *AI threats to national security can be countered through an incident regime* (arXiv No. 2503.19887). arXiv. *https://arxiv.org/abs/2503.19887*.

Prism Scenario Lab. (2025, April 16). *AI governance for resilient global supply chains.* *https://prism.sustainability-directory.com/scenario/ai-governance-for-resilient-global-supply-chains/*

Red Lines. (2025, September). *Call for red lines to prevent unacceptable AI risks.* *https://red-lines.ai/.*

Rodriguez, M., Kim, J., Shah, A., & Patel, R. (2025). *A framework for evaluating emerging cyberattack capabilities in the age of AI* (RAND-style technical report). *https://breached.company/content/files/2025/04/aicyberattacks.pdf.*

Scafaria, L. (Director). (2012). *Seeking a Friend for the End of the World* [Film]. Indian Paintbrush; Mandate Pictures. *https://www.netflix.com/title/70228040*

Stetler, N. (2025). *Reinsuring AI: Energy, agriculture, finance & medicine as precedents for scalable governance of frontier artificial intelligence.* arXiv. *https://arxiv.org/pdf/2504.02127.pdf*

The Economic Times. (2025, November 1). *Xi bats for global AI body to trump US.* The Economic Times. *https://economictimes.indiatimes.com/news/international/world-news/xi-bats-for-global-ai-body-to-trump-us/articleshow/125022016.cms*

Tomei, P. M., Jain, R., & Franklin, M. (2025). *AI governance through markets.* arXiv. *https://arxiv.org/abs/2501.17755*

UK Government. (2023). *AI Safety Summit 2023.* *https://www.gov.uk/government/topical-events/ai-safety-summit-2023*

United Nations General Assembly. (2025). *Resolution A/RES/79/325: Artificial intelligence governance mechanisms.* United Nations. *https://docs.un.org/en/A/RES/79/325*

University of Chicago News. (2023, August 30). *First nuclear reactor, explained.* University of Chicago. *https://news.uchicago.edu/explainer/first-nuclear-reactor-explained*

University of Oxford. (2024, May 21). *World leaders still need to wake up to AI risks, say leading experts ahead of AI safety summit.* *https://www.ox.ac.uk/news/2024-05-21-world-leaders-still-need-wake-ai-risks-say-leading-experts-ahead-ai-safety-summit*

Von Trier, L. (Director). (2011). *Melancholia* [Film]. Zentropa Entertainments. *https://www.netflix.com/title/70184165*

Weidinger, L., Raji, I. D., Wallach, H., Mitchell, M., Wang, A., Salaudeen, O., Isaac, W., & others. (2025). *Toward an evaluation science for generative AI systems* (arXiv No. 2503.05336). arXiv. *https://arxiv.org/abs/2503.05336.*

World Artificial Intelligence Conference High-Level Meeting on Global AI Governance. (2025, July 26). *Global AI Governance Action Plan* [Multistakeholder action plan]. *https://un.china-mission.gov.cn/eng/zgyw/202507/t20250729_11679232.htm.*

World Bank. (2021). *Artificial intelligence in the public sector: Maximizing opportunities, managing risks.* World Bank. *https://documents1.worldbank.org/curated/en/809611616042736565/pdf/Artificial-Intelligence-in-the-Public-Sector-Maximizing-Opportunities-Managing-Risks.pdf*

White House. (2025, July 23). *Winning the AI race: America's AI action plan* [Policy report]. *https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf.*

Yew, R.-J., & Judge, B. (2025). *How "AI safety" is leveraged against regulatory oversight* (arXiv No. 2509.22872). arXiv. *https://arxiv.org/pdf/2509.22872.pdf.*

ANNEXURE 1: AI Risk Calculations

There are many sceptics of "Don't Look Up" scenarios. They believe that a focus on catastrophic risks can obstruct the growth of AI. Since many of the existential risks are in the future, it is difficult to say with confidence if and when they would appear. But if they do become reality, the danger of human extinction or the loss of human autonomy is so fundamental that it would be too late to take any action.

It is therefore necessary to use probabilistic reasoning based on historical precedence and logic. One approach could be to estimate AI risks using the Compton Threshold as a point of reference.

Arthur Compton was the supervisor of Dr Oppenheimer in the Manhattan Project. He had said that he would allow the Trinity Test to go on if the risk of destruction of the world was $\leq 3\times10^{-6}$ or less than 3 in a million. The basis for this calculation was whether the energy production would exceed energy loss.

Physicists worried that the atomic bomb might heat the air so much that nitrogen nuclei in the atmosphere would start fusing, releasing more energy, heating the air further, and creating a chain reaction that could spread across the planet. The key question was not "How likely is this?" but "Can it sustain itself once it starts?"

They reduced the problem to a simple comparison: How fast energy would be produced by such fusion versus how fast energy would be lost through radiation and the rapid expansion and cooling of the hot air. If energy leaked away faster than it was created, the reaction would die out immediately. If energy was created faster than it leaked away, the reaction could grow uncontrollably.

This logic for Risk (R) can be written as:

$$R = \frac{\text{Rate of energy produced}}{\text{Rate of energy lost}}$$

If $R < 1$, a global runaway is physically impossible.
If $R > 1$, catastrophe cannot be ruled out.

Once physicists were confident that R was far below 1, Compton added a policy rule: even then, the chance of global destruction must be less than a few parts per million before proceeding. The lesson is simple: when an experiment could destroy the world, you do not accept ordinary levels of risk - you first prove that a runaway process cannot sustain itself.

Using the Compton Threshold, SFG made a thought experiment to develop three scenarios. We conceived the ideas, the parameters of the Compton Threshold, and then used Chat GPT 5 to make calculations and structure mathematical formulas. In doing so, we want to clarify that our work does not claim that AI risks are directly comparable to nuclear physics in mechanism, only that the logic of excluding self-sustaining runaway processes under extreme uncertainty is transferable. The numerical

bounds and thresholds used here are introduced solely as illustrative reference points for reasoning under deep uncertainty.

Now let's look at three scenarios. While Compton was calculating his threshold for a singular event, if we apply the same logic to AI risks which are an ongoing event, we have to measure the risk on an annual basis.

Scenario I: AI misalignment or malfunctioning in the early warning system of nuclear weapons

$$R = \frac{\text{False escalation impulses per year}}{\text{Human + institutional damping capacity per year}}$$

If R > 1, escalation dynamics can become self-sustaining, as false signals propagate faster than they can be suppressed by human or institutional safeguards.

False escalation impulses come from:

- Sensor error amplified by AI
- Adversarial manipulation
- Human over-trust in AI output.

Let's assume that all kinds of human and machine guardrails are in place.

Let's assign pessimistic but bounded numbers.

- Major nuclear states using AI-assisted warning: 5
- AI-assisted false critical alerts by passing the guardrails per state per year (post-filtering): 0.01

This corresponds to approximately one serious false alert per century per state, consistent with historical precedent.

So:

$$\text{Total AI false alerts/ year} = 5 \times 0.01 = 0.05$$

But some years might have two or more incidents simultaneously.

The relevant global risk metric, however, is the probability that at least one such event occurs anywhere in a given year:

$$P(\text{at least one}) = 1 - (1-0.01)^5 \approx 0.049$$

Expressed in Compton-style units:

$$0.049 = 49{,}000 \text{ per million per year}$$

Compared to Compton's threshold of 3 per million for a one-time experiment, this represents a risk that is over four orders of magnitude larger, and it applies every year, not once.

Scenario II: AI Guardrail Failure Enabling Catastrophic Bio/Chemical Harm

In the Trinity case, the question was whether a physical process could become self-sustaining once triggered.

For advanced AI systems assisting scientific reasoning, the analogous question is:

Can AI-enabled knowledge propagation outpace human, institutional, and biological containment once a dangerous capability is released?

Risk ratio definition

$$R = \frac{\text{Rate of AI-enabled catastrophic capability release}}{\text{Rate of effective global containment and suppression}}$$

- If $R < 1$: Harmful knowledge may appear but cannot propagate widely or fast enough to cause global catastrophe.
- If $R \geq 1$: Capability diffusion outruns containment; catastrophe cannot be ruled out.

Potential sources of AI-enabled catastrophic release include:

- Unexpected jailbreaks or emergent behaviours in frontier models
- Model misuse by moderately skilled actors (not state-level experts)
- Accidental publication or leakage of actionable protocols
- Multi-model toolchains that bypass single-model safeguards.

We assume:

- State-of-the-art guardrails are deployed
- No malicious superintelligence is developed
- Only human-level misuse is amplified by AI assistance.

Pessimistic but bounded numerical assumptions:

- Number of frontier or near-frontier AI platforms worldwide: **~10**
- Annual probability per platform of a *serious guardrail failure* that meaningfully lowers the barrier to catastrophic bio/chem harm: **0.005**
  (≈ one such failure every 200 years per platform; this is very conservative relative to software failure history)

This yields:

$$\text{Expected serious releases/ year} = 10 \times 0.005 = 0.05$$

The probability of at least one serious release anywhere in a given year is:

$$P(\text{at least one}) = 1 - (1 - 0.005)^{10} \approx 0.049$$

So again:

0.049 per year = 49,000 per million per year

Containment comparison (Compton-style insight) in biological systems:

- Response time is measured in weeks to months
- Detection often occurs after spread begins
- Attribution is uncertain
- Replication is exponential by default.

This strongly suggests that once a sufficiently powerful protocol is released, containment rate is unlikely to exceed release rate, implying: $R \gtrsim 1$

Compared to Compton's 3×10^{-6} one-time risk, this is:

- Four orders of magnitude larger
- Recurring annually
- Systemic rather than singular.

Scenario III: Loss of Control via Agentic AI and Self-Amplification

The nitrogen fusion fear was about **runaway physics**.

For advanced AI, the parallel concern is **runaway optimization** (systems that improve, replicate, coordinate, or influence outcomes faster than humans can understand or restrain).

The key question is again not probability of emergence, but sustainability once started.

Risk ratio is defined as:

$$R = \frac{\text{Rate of self-amplification (capability, copies, influence)}}{\text{Rate of human detection, alignment, and shutdown}}$$

- If $R < 1$: Dangerous behaviour is detectable and suppressible.
- If $R \geq 1$: Loss of control becomes plausible.

Without assuming AGI or consciousness, self-amplification may arise from:

- Autonomous agent networks
- Recursive tool use and delegation
- Copying across cloud systems
- Economic and informational influence loops
- Emergent coordination between AI systems.

Pessimistic but bounded numerical assumptions:

- Number of large-scale agentic AI deployments worldwide (next decade): **~20**
- Annual probability per deployment of entering a partially self-amplifying regime that escapes intended oversight: **0.002**
  (≈ one such event every 500 years per deployment; very conservative estimate for complex systems)

This yields:

$$\text{Expected loss-of-control initiations/year} = 20 \times 0.002 = 0.04$$

If we calculate global aggression

$$P(\text{at least one}) = 1 - (1 - 0.002)20 \approx 0.039$$

So:

$$0.039 \text{ per year} = 39{,}000 \text{ per million per year}$$

Human response is limited by:

- Detection lag (often weeks or months)
- Jurisdictional fragmentation
- Economic incentives to delay shutdown
- Difficulty of coordinated global action.

Once multiple AI systems are involved, shutdown becomes non-local and non-instantaneous, strongly suggesting:

$$R \approx 1 \text{ or } R > 1$$

Across all three scenarios including AI involvement in nuclear early warning, AI-enabled release of catastrophic biological or chemical capabilities, and loss of control through self-amplifying AI systems, the same structural result appears. When risk is expressed in Compton-style terms, the relevant ratio R = the rate at which dangerous AI-driven impulses are generated divided by the rate at which human and institutional systems can detect, dampen and suppress them cannot be confidently shown to be far below 1.

Under pessimistic but historically plausible assumptions, the resulting annual probability of at least one globally consequential failure lies in the range of 4×10^{-2}, or roughly 40,000–50,000 per million per year. By comparison, Arthur Compton required confidence that a *one-time* nuclear test carried a risk no greater than 3 per million before proceeding.

The goal of the above framework is not to estimate the true probability of catastrophe, but to determine whether current systems can be confidently shown to lie in a regime where dangerous dynamics decay faster than they propagate.

The critics might argue that the three scenarios are discussed in a sandbox like environment. The real world is complex and risks are correlated. In fact, in a deteriorating geo-political context, the risks identified above will multiply several times. The Compton threshold was tested in a laboratory. The AI systems are run an intense corporate and inter-state competition.

In simple terms, the world is now operating several continuously running systems whose plausible upper-bound risks exceed the tolerance applied to the most dangerous experiment in human history by more than four orders of magnitude. The core lesson is not that catastrophe is inevitable, but that, as with the Trinity test, safety cannot rest on optimism or probability alone. Before accepting ongoing exposure to such risks, it must be demonstrated that runaway dynamics cannot sustain themselves and that the system is firmly in a regime where dangerous processes decay faster than they propagate.

ACKNOWLEDGEMENTS

The following experts reviewed earlier versions of this paper and offered detailed comments. The paper is a result of their input while the overall responsibility for analysis, errors and omissions is solely of the authors.

- Stefan Lofven, Former Prime Minister of Sweden
- Danilo Turk, Former President of Slovenia
- Kyoungjin Choi, Director of Center for AI·Data and Policy at Gachon University, South Korea
- Kyungho (David) Song, Senior Researcher, K-AISI, South Korea
- Jiyeon Cho, Senior Researcher, K-AISI, South Korea
- Guilherme Fitzgibbon Alves Pereira, Special Advisor for Artificial Intelligence Affairs: Department of Science, Technology, Innovation, and Intellectual Property, Ministry of Foreign Affairs, Brazil
- Luis Enrique Urtubey De Cesaris, Director, CEGIA, Brazil
- Fadi Salem, Director of Policy Research Department, Mohammed Bin Rashid School of Government, UAE
- Derrick Swartz, Science Expert, Office of Minister of Department of Science, Technology and Innovation, South Africa
- Nqabekaya Nqandela, Special Advisor, Office of Minister of Science, Technology and Innovation, South Africa
- Mlindi Christian Mashologu, Deputy Director-General, Digital Society and Economy, Department of Communications and Digital Technologies, South Africa
- Mbali Dawn Hlophe, Chairperson, Gauteng e-Government, Research and Development, South Africa
- Brian Tse, Founder and CEO, Concordia AI, China
- Yawen Duan, AI Safety Research Manager, Concordia AI, China
- Jing Shao, Co-Principal Investigator, Center for Safe and Trustworthy AI, Shanghai AI Lab, China
- Jia Xu, Safety and Trustworthy AI Researcher, Shanghai AI Lab, China
- Laura Ihle, Independent, Denmark

In addition, we acknowledge research contribution for Part II on Demand Side Measures by Vivek Kelkar, an independent researcher based in India.

Strategic Foresight Group

Strategic Foresight Group (SFG) is an international think-tank based in Mumbai, India. It has collaborated with governments and national institutions of more than 70 countries, since its inception in 2002. It helps policy makers to anticipate global challenges such as catastrophic wars, transformative technologies, cross-boundary water conflicts and terrorism. It develops new policy concepts and convenes leaders from rival countries to craft collaborative solutions. On the background of the Ukraine war, SFG steered a process of dialogue between five Permanent Members of the UN Security Council to develop norms for governance of AI in the nuclear domain, in partnership with other institutions. Since 2025, it has been involved in developing global norms for AI extreme risks, particularly involving Global Majority countries. The research output and policy prescriptions of SFG have been discussed in the United Nations Security Council, UN Alliance of Civilizations, World Economic Forum at Davos, Interaction Council of former Heads of States, European Parliament, Indian Parliament, UK House of Lords, House of Commons and other institutions.

www.strategicforesight.com

The extreme risks posed by Artificial Intelligence to the continued survival of human civilisation has been a subject of worldwide discussions. World leaders, including UN Secretary General, Pope Leo XIV, Presidents of China and the United States, have voiced concerns about various implications about extreme AI risks. Some countries have introduced legislations, guidelines or industry frameworks to contain these risks. This report explores convergence between various subnational, national and regional solutions presented in the Global North, as well as the Global South. It proposes concrete measures that the suppliers and consumers of the new technology can implement. It also introduces a menu of a few integrated global proposals for discussion in diplomatic fora and elsewhere. It is expected that The Essential Convergence will move worldwide discourse on extreme AI risks from expression of concerns to the exploration of concrete and collaborative solutions.

ISBN: 978-81-88262-37-3