Governing the Frontier: Priorities for International Cooperation on Advanced AI

Centre pour la Sécurité de l'IA

Table of Contents

Authors: *Charles Martinet, Su Zeynep Cizem, Jared Perlo, Jérôme Barbier*

Advisors: *Arthur Grimonpont, Camille Truchot, Ketana Krishna, Jacob Schaal, Alvin Wong, Clément Loup-Forest, Quentin Flament, Inès Belhadj, Felicia Khor, Anthony Chi*

About Us

CeSIA is a Paris-based non-profit organization dedicated to understanding, preventing, and mitigating the risks associated with artificial intelligence. Our core mission is to foster an AI safety culture within France and Europe by promoting technical, policy, and regulatory approaches to address critical AI challenges.

Our strategy relies on four main activities:

1. Institutional Advocacy: We engage with key national (French), European, and international institutions, policymakers, regulatory bodies, and industry actors to ensure AI risks are effectively considered in public policy and regulatory frameworks.

2. Education and Training: We develop and deliver specialized educational content, including university-level courses, intensive training programs, and accessible learning resources, to build AI safety expertise within the academic community and the broader ecosystem.

3. Research and Development: Our research efforts focus on creating foundational resources, developing robust evaluation methods and tools for AI systems, and advancing the theoretical understanding needed for effective AI oversight and safety.

4. Public Awareness and Engagement: We foster broader understanding and informed discussion about AI safety by organizing expert events, publishing analyses and commentary, collaborating on diverse outreach initiatives, and creating platforms for interactive engagement.

Foreword

Artificial intelligence (AI) is one of the most pressing issues of our time, and its advancing capabilities stand to revolutionize almost every aspect of our lives. However, current laws, regulations, and policy frameworks, especially on the international level, are insufficient to address and shape this frontier technology's design, deployment, and impacts. The scale and complexity of this governance challenge will increase in the coming years as AI systems become more powerful and more entangled into our society.

Of the governance initiatives that do exist, many current international mechanisms are too slow, fragmented, and incomplete to address the risks posed by increasingly powerful AI systems. These international systems often overlap in mandate, allow for critical gaps on key issues, or lack the enforcement mechanisms necessary for effective oversight. As a result, the international community must strengthen current international AI governance mechanisms and establish new structures where required. This work must begin *now*, before the pace of AI development outstrips our collective capacity for governance.

This report proposes a focused agenda to address current shortcomings and strengthen global AI governance. We identify **nine priority areas for action**, structured around **Four Pillars for Effective Global AI Governance**. Some of our recommendations call for near-term actions achievable with limited coordination; others are ambitious and will require sustained political and technical cooperation. Together, they offer a roadmap for ensuring that AI development remains aligned with societal interests, and that safety measures are in place before, not after, a crisis emerges.

Executive Summary

AI development has entered a period of explosive growth. While leading companies announce breakthrough after breakthrough, mechanisms to govern AI are either nascent or otherwise unable to sufficiently address growing risks and harms. Without swift action to establish robust global oversight, we risk losing the opportunity to shape how these powerful technologies evolve.

Structural Gaps

The international community currently faces significant obstacles in governing advanced AI systems. Scientific understanding of AI is still nascent and fragmented across institutions and borders, and mechanisms for ensuring AI safety have not kept pace with rapid technological progress. The absence of global standards leaves society ill-equipped to manage shared risks, undermining trust in the development and deployment of frontier AI. These structural issues are compounded by systemic challenges:

Fragmented Governance Efforts — No global mechanism exists to manage AI risks across borders. Current governance efforts are fragmented across national and multilateral initiatives with overlapping mandates and limited cooperation. This undermines collective risk mitigation and allows regulatory arbitrage.

Competing Interests — Competition among leading AI powers has largely sidelined shared safety commitments. Recent developments, such as the weakening of safety language at the 2025 Paris AI Action Summit, reflect the difficulty of aligning national interests. Meanwhile, the global governance landscape remains divided between voluntary and binding approaches, complicating efforts to ensure accountability.

Participation and Capacity Gaps — Countries in the Global South are underrepresented in international processes and often lack the technical infrastructure and policy expertise to engage meaningfully. This limits the legitimacy and effectiveness of global governance efforts.

Underdeveloped Safety Research — Technical methods for red-teaming, interpretability, and systemic risk analysis are not yet standardized nor widely adopted. AI safety research remains underfunded, with only a small share of AI publications and public investments dedicated to this area.

Differing Cultural and Value Approaches — Misalignments between AI safety communities and traditional diplomatic institutions have slowed policy uptake. Scientific assessments of catastrophic risks often fail to translate into actionable frameworks, and existential concerns are frequently dismissed as abstract or alarmist.

Recent Progress

Despite these challenges, we now face a critical window to shape global AI governance. AI capabilities have not yet outpaced our ability to govern them, and several trends create favorable conditions for coordinated international action:

Voluntary Safety Frameworks — Recent commitments by frontier AI companies and states as part of the Seoul Summit, the G7 Hiroshima Process, and through regional partnerships indicate growing momentum for the adoption of high-level AI safety principles and norms. This momentum must now be operationalized into specific technical and procedural cross-border rules and mechanisms.

Growing Expert Consensus on AI Risks – AI researchers increasingly warn of potentially catastrophic outcomes and have called for urgent regulatory action. Expert consensus papers and survey data show broad support for new oversight institutions, mandatory risk assessments, and scalable safety infrastructure.

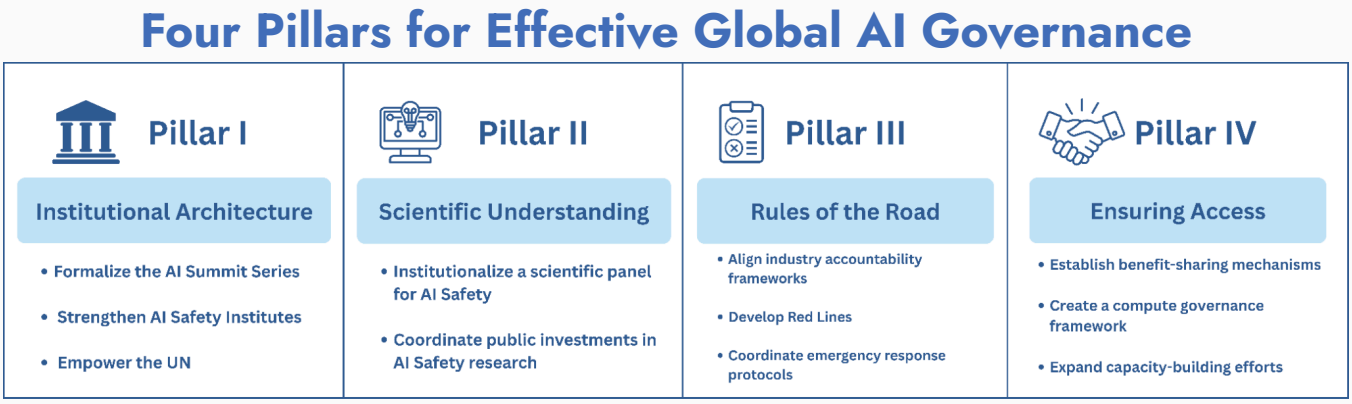
Concentrated Developer Landscape – The current AI landscape is still relatively centralized, with a small number of companies and governments leading frontier development. This makes coordination more feasible than it may be in the future.

Rising Public Interest and Awareness – Citizens and civil society groups are engaging with AI governance more actively than ever. Public consultations have revealed shared global priorities around inclusivity, transparency, labor protection, and equitable access to AI benefits.

Increased Representation of Global Stakeholders – Though international representation is still highly inadequate, many countries in the Global South are now advancing national AI strategies and contributing to regional AI frameworks. Initiatives like the African Union's AI Strategy and the Global Digital Compact show how emerging voices are shaping the global conversation.

Potential for Safety Investment Growth – Although investment in AI innovation is expanding rapidly, the proportion allocated to safety research remains disproportionately low. This disparity presents an opportunity to channel public and private resources into institutions, tools, and talent focused on risk mitigation and technical assurance.

These converging factors point to the need for an integrated and urgent global response. Our framework identifies four essential pillars for building effective global AI governance and offers a practical agenda for addressing the risks and opportunities posed by advanced AI.



Pillar I: Institutional Architecture

Formalizing the AI Summit Series

The AI Summit Series, launched at Bletchley Park in 2023 and continuing through Seoul and Paris, represents the primary high-level political forum for AI governance but faces critical structural challenges. While successful in establishing initial safety commitments and creating national AI Safety Institutes, the series struggles with maintaining legitimacy and effectiveness as it matures. The expansion of the series' scope and membership has created fundamental tensions between effectiveness and inclusivity, exemplified by the failure to achieve meaningful progress on safety issues at the Paris AI Action Summit. The series currently lacks

essential governance mechanisms, including institutional continuity, accountability frameworks, and clear scope. Without better-defined mandates and alignment with complementary forums, there is significant risk of duplication or drift from its core mission of addressing frontier AI safety.

Key recommendation: Turn the AI Summit Series into a sustained, structured, and inclusive international process.

Create a focused track for advanced AI safety coordination among countries with advanced AI capabilities and major developers, alongside a broader track addressing societal impacts. Each track would produce concrete deliverables through annual high-level summits, with a technical-ministerial meeting held at the six-month mark between each summit. This regular cadence would ensure consistent high-level political engagement while enabling mid-year review of commitments and implementation progress.

Strengthening The Network of AI Safety Institutes

National AI Safety Institutes (AISIs) have emerged as central technical bodies for frontier model evaluations, with a new International Network (INASI) launched in 2024. These institutes work at the technical level to evaluate frontier systems and translate safety insights into actionable policy inputs. While showing promise for coordinated safety assessment, the network requires stronger institutional frameworks and standardized evaluation methodologies to prevent fragmentation and enable consistent oversight across jurisdictions. A coordinated network offers a promising solution to align technical evaluations while supporting under-resourced regions, but this requires robust information-sharing mechanisms and an inclusive institutional structure.

Key recommendation: AI Safety Institutes should collaborate to create best practices in AI evaluations, standards, and risk thresholds.

Develop common testing protocols for models exceeding defined capability thresholds, establish third-party evaluation guidelines, and launch a coordinated evaluation program by mid-2026. Build international AI safety standards by 2027 covering risk assessment, alignment evaluations, and safety engineering practices, supported by a secure shared repository of evaluation results.

Solidifying The Role of the United Nations

The UN provides unmatched legitimacy for global AI governance but faces significant challenges in bureaucratic agility and technical capacity. Recent initiatives such as the Global Digital Compact and UN expert panels demonstrate growing engagement with AI governance, though the UN system currently lacks clear mandates and dedicated institutional capacity for addressing advanced AI risks. Translating recent resolutions on the potential large-scale risks of advanced AI systems into effective oversight mechanisms remains a

challenge. The UN must balance its unique convening power with the need for nimble responses to rapidly evolving AI capabilities.

Key recommendation: Negotiate a binding international framework convention for AI safety by 2027

Following comprehensive legal analysis of existing international law's applicability to AI systems, develop a convention establishing core principles and processes for international AI governance. Create subsequent protocols defining state obligations for frontier AI development, risk management, and prevention of catastrophic misuse.

Pillar II: Scientific Understanding

Increase Scientific Assessment of Advanced AI

The international community lacks sustained mechanisms for scientific assessment of AI progress and risks, hindering evidence-based policymaking. While the first International AI Safety Report marked an important milestone in building scientific consensus, the current ad-hoc approach is insufficient for the rapid pace of AI development. The report lacks a permanent host institution and sustainable funding, limiting its ability to provide ongoing, technically grounded analysis. Scientific assessment processes also face several key challenges, including reconciling the speed of AI development with traditional peer review processes, addressing divergence between scientific findings and political priorities, maintaining a focused yet inclusive mandate, and integrating non-traditional sources of evidence such as preprints and industry evaluations. More permanent and independent scientific assessment bodies are needed to build sustained consensus on AI risks and capabilities.

Key recommendation: Institutionalize the International AI Safety Report process within the UN Independent Scientific Panel on AI

Strike a balance between the political legitimacy enabled by UN and member state support and the scientific legitimacy created by a robust process for producing evidence, while remaining transparent, inclusive, and globally representative.

Research Funding and Coordination

AI safety research is severely fragmented and underfunded compared to capabilities research, creating critical gaps in our understanding of AI risks and mitigations. The field lacks coordinated research agendas, shared infrastructure, and sustainable funding mechanisms, leading to duplication of efforts and missed opportunities for collaboration. Current funding structures and incentives often prioritize short-term capabilities advances over longer-term safety considerations. International collaboration faces complex challenges in balancing open science with security considerations, particularly around access to advanced models and sensitive research

findings. Without better coordination and resources, safety research may continue to lag behind capabilities development, increasing societal vulnerabilities to AI risks.

Key recommendation: Introduce a funding paradigm for AI safety research, requiring qualified AI companies to contribute to AI safety research through a structured funding mechanism

Implement a tiered funding system requiring companies developing advanced AI (e.g. >10^25 FLOP) to contribute a percentage of R&D investments. Allow both monetary and non-monetary contributions (compute, data, expertise), with 15-25% allocated to a global fund for multinational projects and the remainder distributed through national funds.

Pillar III: Rules of the Road

Industry Accountability

Current AI governance relies heavily on voluntary industry commitments without robust verification mechanisms, creating significant oversight gaps. While the accountability ecosystem has evolved from early ethics principles toward more specific technical commitments about risk assessment and deployment safeguards, the absence of standardized reporting requirements and independent verification mechanisms undermines trust in advanced AI systems. More systematic approaches to monitoring commitments and assessing risks are needed, including standardized reporting frameworks, third-party evaluation mechanisms, and clear consequences for non-compliance. Such measures would help institutionalize developer accountability at key stages of the AI lifecycle.

Key recommendation: Align on an international system for AI companies' public reporting on risk management and responsible AI

Create a standardized reporting template compatible with existing commitments and frameworks, structured around key performance indicators and qualitative elements. Require regular updates and immediate disclosure of serious incidents.

Risk Thresholds and Red-Lines

The international community currently lacks clear thresholds and red lines for AI development to define unacceptable levels of risks from AI, creating significant uncertainty and hazard. While various regulatory frameworks propose risk tiers, there is no common understanding of thresholds that should trigger specific oversight measures or capabilities that should be prohibited. Without clear thresholds and enforceable red lines, there is a growing risk that developers will cross into dangerous territory, whether through deployment of deceptive agents, autonomous bio-threat design, or other catastrophic misuse. Early warning systems and clear communication protocols will be essential for effective implementation.

Key recommendation: Establish international consensus for categories of unacceptable risks (or "red lines") to prevent large-scale harm

Define specific AI capabilities requiring strict limitations, such as autonomous cyberattacks or self-replication, with regular reviews to account for emerging threats. Establish trigger points for regulatory intervention when red lines are crossed, building on existing international declarations and agreements.

Pillar IV: Ensuring Access

Benefit Sharing

Current AI infrastructure, development, and deployment is largely concentrated in a select few countries, primarily in the US and China. As a result, computing resources, talent networks, and potential economic dividends largely elude most of the world's countries and population. This gap in benefit distribution will likely be exacerbated over time as AI development increasingly benefits the countries and actors that develop the technology compared to downstream users. Current efforts remain very preliminary and insufficient to ensure adequate, equitable, and secure access to AI and its potential benefits. The potential for AI to cause immense shifts in labor markets is also woefully understudied and under-addressed. The possibility that AI could lead to widespread unemployment or labor redistribution demands much more attention from governments and international organizations. International frameworks must ensure more equitable distribution of AI infrastructure, talent development opportunities, and economic returns while maintaining appropriate safety standards.

Key recommendation: Establish a global financial AI benefits redistribution mechanism with clear allocation formulas

Create a framework requiring companies exceeding specific revenue thresholds from AI products to contribute to a global fund. For example, this might take shape as a fee on frontier AI revenue, with progressive rates from 1-3% of AI-derived profits. This windfall could then allocate resources to universal basic income programs, AI education programs, and public-interest AI applications.

Export Controls and Compute Governance

Computing power is critical to every step within the development and use of AI. As a result, AI-relevant compute is an especially effective entry-point or lever for controlling the development of advanced AI systems. While compute access can be leveraged for oversight and control of AI development, overly restrictive controls risk hampering economic development in many regions. A balanced approach is needed to enable legitimate development while preventing misuse, requiring international coordination mechanisms that can differentiate between legitimate development needs and security risks.

Key recommendation: Provide compute resources only to countries with robust safety and security mechanisms

Establish a conditional compute access framework for international AI development. Provide computing resources and chip access to countries that implement robust safety and security protocols, including physical safeguards for data centers and cybersecurity measures for cloud computing. Leading compute-capable nations should coordinate to ensure consistent safety standards, while an independent oversight initiative should verify compliance through regular inspections and audits.

1. Introduction

1.1 The State of International Governance

Over the past several years, leaders from academia, civil society, industry, international organizations, and individual nations have made significant progress in establishing guidelines and regulatory frameworks to ensure safe AI development and deployment. The international AI governance landscape now consists of a complex web of initiatives operating at different levels with varying mandates. This includes:

Series of Global AI Summits — Perhaps the best-known international AI governance effort, the Bletchley Park Process and the ensuing AI Summit Series has emerged as a unique platform for global coordination on AI safety. Launched by the United Kingdom, it brought catastrophic risks from advanced AI into high-level political discussions for the first time, elevating AI safety from a niche concern to an international priority. The original Bletchley Park Summit in 2023 produced the Bletchley Declaration, a landmark agreement signed by 28 countries and the European Union, and established a shared commitment to assess and mitigate AI risks. This outcome held added significance as even the United States and China, engaged in a growing geostrategic competition, were interested in global cooperation for the safe and responsible development of AI[1]. Building on this foundation, the Seoul Summit in May 2024 expanded discussions to include not only safety but also innovation and inclusivity, reflecting the increasing intersection of AI policy with economic and social priorities. The Paris AI Action Summit, held in February 2025, sought to broaden the global AI conversation by shifting away from a primary focus on AI safety. This shift has faced criticism for deprioritizing discussions on catastrophic risks and derailing the summits from their original focus. While the AI summit process continues to evolve and has clear limitations, its early stages brought AI safety to the forefront of international governance debates above and beyond previous multilateral efforts. However, with the broadening of the summits' scope over time, the role of AI safety in future summits remains to be seen.

AI Safety Institutes (AISIs) — A major outcome of the series of global AI Summits has been the establishment and expansion of AI Safety Institutes (AISIs) across several countries. These institutes act as hubs for public AI safety expertise and evaluation of advanced AI systems, particularly focusing on frontier models[2]. The United States and United Kingdom have led this initiative, formalizing cooperation through memorandums of understanding between their respective institutes. Several other countries and regions have announced plans for their own institutes, strengthening the global infrastructure for AI safety and governance. As of January 2025, these include Australia[3], Canada, the European Union, France, Japan, the Republic of Korea and Singapore.

---

[1] The United States and China mutually agreed to adopt the first two UN resolutions on AI, at the 78th session of the United Nations General Assembly, one led by each power. They further established bilateral mechanisms to share insights about their mutual perceptions of threats and critical risks resulting from the widespread diffusion of the technology.

[2] We define frontier (AI) models as *general-purpose models that outperform all other models that have been widely deployed for at least 12 months, as scored on a range of conventional performance benchmarks or high-risk capability assessments.*

[3] Australia is the only signatory to the declaration that has yet to meet its commitments.

International Network of AI Safety Institutes (INASI) — On November 21, 2024, representatives from these countries met in San Francisco to launch the **International Network of AI Safety Institutes (INASI).** In its inaugural convening, the network issued a joint statement outlining foundational principles for the assessment of advanced AI risks, emphasizing transparency, methodological rigor, and real-world applicability. The convening also saw announcements of new public and philanthropic funding, including over $11 million in support for international AI safety research. The creation of a new Testing Risks of AI for National Security (TRAINS) Taskforce under the U.S. AISI to address national security implications of AI. INASI intends to provide a structured forum for technical coordination, shared model evaluation, and joint safety research. While national AI safety institutes retain jurisdictional authority, INASI enables cross-border information sharing and standard-setting, helping to build a more interoperable global infrastructure for AI governance. However, the political uncertainty surrounding the U.S. AISI's future (as of this report's publication) brings into question the commitments and institutional resilience across participating countries.

Selected Other Global and Multilateral Initiatives — Several other initiatives also play important roles in the global AI governance landscape. The **G7 Hiroshima AI Process (HAIP)** and the involved nations hold significant sway in global rulemaking due to their economic, regulatory, and technological leadership. In 2023, the G7 accounted for approximately a quarter of global GDP, and most frontier AI companies are based in one of its member nations. The HAIP has focused on establishing interoperable rules for advanced AI systems, reducing compliance costs, and facilitating global innovation. **The United Nations (UN)** system has engaged in AI governance through several avenues, including UNESCO's Recommendation on the Ethics of AI, the High-level Advisory Body on AI, which convened 39 experts from 33 countries to formulate recommendations, and the recent transformation of the Tech Envoy's Office into the Office for Digital and Emerging Technologies intended to support implementation of the Global Digital Compact, including its AI commitments. The UN is now developing a two-track Panel and Dialogue on AI though the exact form of this effort remains unclear. Nonetheless, the Panel and Dialogue have immense potential to transform and centralize international AI governance. In addition, **the Organization for Economic Co-operation and Development (OECD)**'s AI Principles adopted in 2019 emphasize transparency, accountability, and respect for human rights, providing a foundation for many regional governance efforts. For a summary of selected regional approaches, see *Annex*.

1.2 The State of AI

Current AI systems already demonstrate capabilities that raise significant security and safety concerns, while the barriers to developing and deploying powerful AI continue to decrease. This combination of advancing capabilities and increasing accessibility demands immediate attention from policymakers to prevent potential adverse outcomes.

Current AI Capabilities — AI systems have already demonstrated remarkable abilities across a broad range of domains, often matching or exceeding human performance in complex tasks. Leading models surpass 93% of programmers in professional coding competitions, and demonstrate greater proficiency than most law students on bar examinations. On some metrics, AI has also demonstrated creative capabilities, with systems winning art competitions and displaying superior performance on standardized creativity assessments, where they outperform 99% of humans. AI systems have become more mainstream in recent years due to their sophisticated language representation and generation abilities, enabling them to translate between languages with near-human accuracy and produce highly convincing text across various styles and contexts.

Scientific Breakthrough Capabilities — Moreover, certain AI systems exhibit capabilities entirely absent in humans, such as accurately predicting protein structures—recognized as a breakthrough achievement by the 2024 Chemistry Nobel Prize—and generating highly realistic synthetic images and videos from scratch. AI systems are increasingly becoming capable of not just *analyzing* but also *generating* novel scientific insights, potentially accelerating research timelines significantly. Recent research has demonstrated further expansion of AI capabilities into scientific discovery, generating research hypotheses about bacterial transformation and antimicrobial resistance mechanisms identical to those developed by human scientists, but in a fraction of the time. Current models can also solve novel mathematical problems and autonomously design and optimize algorithms, suggesting potential for self-improvement capabilities that could lead to even more advanced systems. This self-improving potential of AI technology is a great challenge for governance, as systems may enhance their capabilities faster than regulatory frameworks can adapt.

Security and Dual-Use Risks — Such heightened capabilities bring significant risks, particularly related to cybersecurity, bio-hazards, and other dual-use applications. In cybersecurity, the same characteristics that make AI valuable for defense (such as speed, scalability, and adaptability) can also create vulnerabilities where AI could potentially outpace defensive measures. AI models can now match the capabilities of top human penetration testers, a significant milestone with complex implications. These systems can autonomously identify and exploit vulnerabilities at unprecedented speed and scale, overwhelming traditional defense mechanisms. AI-powered malware can adapt to evade detection, while AI-generated phishing attacks can be more sophisticated and harder to detect than their traditional counterparts. In biology, the same model able to design a benign viral vector to deliver gene therapy could be used to design a pathogenic (and potentially lethal) virus capable of evading vaccine-induced immunity, sufficient to trigger a pandemic. While evaluations show that models are not yet able to cause catastrophes, it seems that the next generation of models might cross several safety thresholds. For example, Claude 3.7 Sonnet is approaching unacceptable risks for bioweapon development capabilities. And unlike many traditional weapons of mass-destruction, AI-enabled hazards could be created by both state and non-state actors. AI can significantly lower the threshold for bad actors to design malicious uses and actually act on those ideas (e.g. building bombs). With even more increased capabilities, frontier AI's increased capabilities will present significant new risks.

Accelerating Development Factors — The progression of AI capabilities is accelerating due to several technical and economic factors. For example, the emergence of the test-time compute paradigm has revealed previously hidden capabilities in existing systems. This allows AI to "think" before answering, leading to significantly improved performance without the need for expensive retraining.

As computational costs continue to decrease following historical trends in processing power and storage costs, the financial barriers to developing and deploying sophisticated AI systems are rapidly diminishing. The diffusion[4] of research and development, while beneficial for innovation, also means that powerful AI tools are proliferating across industries. While broader access to advanced AI systems could accelerate innovation, widespread diffusion also means that AI systems could easily be implemented without adequate safety measures or oversight. Organizations with limited AI safety oversight may deploy increasingly capable systems without fully understanding or mitigating their potential risks.

Growing Expert Consensus —While precise predictions about the pace of AI development are challenging, there is growing consensus among leading experts regarding the trajectory toward advanced capabilities. All

[4] *Diffusion* refers to the spread or distribution of capabilities and knowledge across various organizations, researchers, and developers. This includes the public release of AI models, training methodologies, and research findings.

three recipients of the 2018 Turing Award[5] have indicated that artificial general intelligence (AGI) is increasingly likely to appear in the coming years—though the definition of AGI varies between actors[6]—and leading AI labs have projected varying timelines for the emergence of human-level or transformative AI systems. Different actors have estimated these supercharged systems' arrival anywhere from *the next year* to within the *next decade*. Though timelines are hard to estimate, there is growing consensus that we are approaching a critical threshold in AI development, with systems exceeding human-level capabilities across all domains.

The combination of rapidly advancing capabilities, decreasing development costs, and significant potential for transformation creates an urgent need for policy frameworks that can effectively govern this technology. Without proper oversight and safety measures, the continued advancement and proliferation of AI systems could pose substantial risks to cybersecurity, public safety, and global stability.

[5] Yoshua Bengio, Geoffrey Hinton, and Yann LeCun

[6] For example, Google defines AGI as "the hypothetical intelligence of a machine that possesses the ability to understand or learn any intellectual task that a human being can," while OpenAI defines AGI as "highly autonomous systems that outperform humans at most economically valuable work."

2. Structural Gaps

The international community currently faces significant barriers to governing advanced AI systems. Scientific understanding remains nascent and fragmented across institutions and borders, while mechanisms for developing and verifying safety measures have not kept pace with rapid technological advancement. The absence of global standards leaves us ill-equipped to address risks and safety challenges, severely undermining trust in advanced AI. These challenges are further complicated by tensions between the need for rapid decision-making in response to advancing AI capabilities, and the imperative for inclusive processes that ensure broad international buy-in.

2.1 Gaps in International Governance

A critical lack of global mechanisms to address AI risks — Risks from AI systems are not confined to the country in which they are developed. When a powerful AI system developed in one country is deployed, its impacts can be felt worldwide. Despite recent progress, significant gaps remain in the international governance of AI. Existing frameworks are either voluntary or geographically limited, enabling regulatory arbitrage[7] and undermining individual nations' efforts to govern AI effectively.

Institutional Fragmentation — The current landscape also suffers from multiple organizations and initiatives operating in parallel with often-overlapping mandates and limited coordination. While the Bletchley Park process has introduced some political incentive for coherence, concerted effort and sustained political will are necessary to convert the process into a more comprehensive governance framework. This fragmentation enables both states and private actors to selectively engage with governance frameworks. Without coherent institutional architecture and binding commitments, the international community risks creating an organized hypocrisy, i.e. formal structures that project regulatory legitimacy while enabling divergent national paths with minimal substantive coordination.

A lack of representation of diverse stakeholders — Many developing nations, particularly from the Global South, are underrepresented in key initiatives, raising questions of legitimacy and potentially leading to governance frameworks that fail to account for their needs. These countries often lack the technical expertise and resources to participate effectively, exacerbating technological disparities and limiting the effectiveness of governance efforts.

2.2 Challenges for International Governance

The rapid pace of AI development — Traditional governance mechanisms require time to build consensus and lengthy negotiations when adopting binding international instruments. More innovative approaches such as the Intergovernmental Panel on Climate Change (IPCC) model for scientific assessment still struggle to match the speed and scale of AI advancements, highlighting the need for more agile and responsive processes.

[7] Where companies or developers choose to operate in jurisdictions with weaker or less enforceable AI regulations.

Balancing the need for rapid decision-making with the need for inclusive processes — While the speed of AI innovation demands quick responses, meaningful international cooperation requires time for consultation, negotiation, and consensus-building among multiple actors with conflicting interests, values and worldviews. The Bletchley Park process has attempted to address this through a series of summits dedicated to AI Safety at high political level, but the challenge of maintaining both momentum and inclusivity while getting to concrete measures in the short term remains.

Voluntary and mandatory approaches to governance — The European Union's binding regulatory approach through the AI Act contrasts sharply with the United States' reliance on voluntary commitments from industry. Voluntary frameworks can be more flexible and adaptable to rapid technological change, while mandatory regulations offer direct enforcement mechanisms and greater accountability. Finding the right balance between these approaches remains a significant challenge. At the international level, initiatives remain largely non-binding "soft law" arrangements that lack verification mechanisms, enforcement provisions, or reporting on implementation. Most initiatives avoid establishing concrete rules altogether, instead focusing on high-level principles and voluntary standards despite growing consensus on both the need for international rules and their potential content.

The perceived tradeoff between encouraging innovation and ensuring safety — Overly restrictive regulations could stifle beneficial AI development and put jurisdictions at a competitive disadvantage. However, insufficient oversight could lead to unsafe systems or misuse of AI technologies. Weighing the benefits of transparency and distributed innovation against the potential security risks of uncontrolled development presents a major regulatory challenge, apparent in the debates around **open-source** and **open-weights**. Open-sourcing offers advantages such as enabling external oversight, accelerating progress, and decentralizing control over AI development and use. However, it also presents a growing potential for misuse and unintended consequences. Open-sourcing has historically provided substantial net benefits for most software and AI development processes. On the other hand, for highly capable foundation models likely to be developed in the near future, open-sourcing the weights may pose sufficiently extreme risks that outweigh the benefits.

The concentration of AI development capabilities in a small number of nations — Countries with strong digital infrastructure, easy access to large amounts of both public and private funding, a developed digital economy with very large companies based in their jurisdiction, and large specialized workforces currently lead in developing advanced AI systems. At the same time, many other nations lack the resources to participate meaningfully in their development. This imbalance raises important questions about ensuring equitable governance frameworks while addressing the responsibilities that come with leading AI development. Countries from the Global South, in their diversity, also emphasize the need to focus efforts on democratizing access to AI technologies and maximizing their economic development benefits.

Cultural and value differences in establishing global governance frameworks — Because state delegations and actors from the larger international governance ecosystem have approached AI challenges with traditional concepts and mechanisms from the diplomatic field, they have often focused on short-term challenges and procedural issues rather than grasping with the radical novelty of AI technologies. Contributing to this situation, the AI safety community has largely struggled to attract attention from non-specialized decision-makers regarding the high-level risks posed by unsafe AI systems, underestimating the importance of established practices and taking into account other policy debates occurring for other digital or emerging technologies. At the same time, existing international governance systems offer underutilized benefits of

structure and negotiating expertise to AI safety advocates. Finding the right balance between a renewal of the existing diplomatic playbook and the necessary integration of any AI governance framework into the existing rules-based international order will be critical for the efficiency and sustainability of any agreed framework.

The discussion of **existential risks** in international policy forums is an important example of a fundamental disconnect between different communities of practice. Scientific assessments of potentially catastrophic risks from advanced AI systems often fail to gain traction in policy circles in part because they are communicated in ways that appear disconnected from established institutional frameworks and diplomatic conventions. **This communication gap results in substantive technical concerns being dismissed or marginalized** in favor of more familiar policy challenges, even when the potential consequences are significantly more severe. The resulting governance approaches may address immediate regulatory needs while leaving critical long-term safety inadequately addressed.

The changing landscape of institutional momentum — As mentioned in Section 1, the Paris AI Action Summit marked a turning point in the international discourse on AI governance. Unlike the Bletchley (2023) and Seoul (2024) summits which produced widely supported declarations, the Paris gathering concluded instead with a statement that notably lacked support from the United States and the United Kingdom. On this occasion, the new US administration strongly advocated prioritizing *innovation* over *regulation*, shifting focus away from safety toward economic opportunities and competitive advantage. This stance followed the repeal of the US Executive Order on AI Safety (EO 14110) in the first week of the new administration, which created significant uncertainty regarding the future role and authority of the US AI Safety Institute. Ahead of the Paris Summit, the UK AI Safety Institute underwent a rebranding to become the UK AI Security Institute, a subtle yet significant shift in terminology that signals a broader distance from safety. Although the Paris Summit's concluding statement acknowledged existing international AI initiatives from various bodies, the statement gave only passing mention to AI safety[8], merely *noting* rather than endorsing or encouraging the voluntary safety commitments that had been central to previous summits.

While the earlier commitments from Seoul, Hiroshima, and other initiatives present an opportunity to formalize efforts into international technical standards and institutionalized governance mechanisms, recent discourse has introduced a narrative that frames innovation and safety as opposing rather than complementary forces. Policymakers now face the complex task of navigating beyond this constructed dichotomy to develop consistent frameworks that recognize safety as foundational to meaningful innovation.

Rising geopolitical tensions and the risks of an "AI arms race" — In an increasingly tense geopolitical environment, leading the race to next-generation AI systems has become a critical strategic consideration for many great and middle powers. Such political investment, sometimes at the expense of safety efforts, has prompted concerns about an "AI arms race" that could undermine international peace and security in similar ways as conventional and non-conventional proliferation did in the past. The successful adoption of self-limitation measures and the establishment of efficient communication channels between leading powers, perhaps regarding new (unreleased) model capabilities or dangerous applications of AI, will be key to steer AI development in a positive direction. Absent this cooperation, we risk a dangerous race towards capabilities whose limits and implications are still unclear.

[8] For example: the UN, UNESCO, African Union, OECD, Council of Europe, EU, G7, and G20, and referenced upcoming events such as the Kigali Summit, the 3rd Global Forum on Ethics of AI in Thailand, and the 2025 World AI Conference.

3. Recent Progress

While AI systems have already achieved impressive capabilities, many experts believe we are at a unique moment in which **meaningful governance frameworks can still be established** before systems become so advanced they will be difficult to regulate and control. Timing is crucial for two reasons: first, current AI systems are powerful enough to demonstrate concrete risks that motivate action but are not so advanced that control becomes impractical. Second, frontier model development remains relatively concentrated amongst a handful of companies. This concentration makes coordination and oversight more feasible than it might be in a more fragmented future landscape. Below, we highlight several converging factors that offer an opportunity to build sustainable frameworks that address some of the challenges of AI governance.

3.1 Voluntary Safety Frameworks

Significant progress has been made toward aligning global efforts regarding the safe development of advanced AI systems. For example, leading frontier AI companies made voluntary AI safety commitments at the 2024 AI Summit in Seoul and states made encouraging political declarations during the G7 Hiroshima AI Process. Such agreements outline risk-based approaches to AI governance, signaling growing acknowledgment that safety, transparency, and accountability are essential to the technology's long-term viability. The G7 Hiroshima Process's Code of Conduct introduced post-deployment vulnerability monitoring, information sharing between industry and governments, and efforts to secure AI systems against physical, cyber, and insider threats as key components of AI risk management frameworks. At the Seoul Summit, commitments included assessing and mitigating risks across the AI model lifecycle, monitoring and enforcing risk thresholds, improving internal risk assessment capabilities, and fostering public transparency. These frameworks also call for **international collaboration in developing technical standards** and **conducting research on societal impacts** of advanced AI. The United Nations resolution to promote safe, secure, and trustworthy AI adopted by the General Assembly in March 2024, followed by the adoption of the Global Digital Compact in September 2024, demonstrate consensus on the need to address AI risks. The Council of Europe Framework Convention on AI and Human Rights was **the first legally binding international treaty** for aligning AI with human rights, democracy, and the rule of law. It has already been signed by 14 countries, in addition to the European Union[9] and is endorsed by the International Bar Association, indicating significant legal and expert support.

In August 2024, the US AI Safety Institute announced agreements with Anthropic and OpenAI to establish formal collaborations on safety research, as well as in model testing and evaluation. The institute also held technical dialogues with the EU's AI Office on best practices and risks. These agreements formalized partnerships for model evaluations, pre-release safety testing, and independent feedback on potential safety improvements. By granting regulators access to cutting-edge AI models, these collaborations signalled that companies are willing to engage in voluntary external oversight and technical validation to mitigate advanced AI risks[10].

[9] This number is accurate as of April 2025.

[10] It should be acknowledged that frontier model companies engage with governance efforts in recognition that regulation is essential for maintaining public trust. While these voluntary measures mark progress in the right direction, they come

3.2 Growing Expert Consensus on Risks

Scientific experts are increasingly unified in warnings about AI risks and the need for stronger governance. Even on more contentious issues like the potential for existential risks from advanced AI systems, there is growing recognition that these concerns warrant serious consideration and coordinated response. In the largest survey of AI researchers to date, between 38% and 51% of respondents estimated that there is at least a 10% chance that advanced AI will lead to outcomes as bad as human extinction[11].

Ahead of the May 2024 AI Seoul Summit, 25 of the world's leading AI scientists[12] published an expert consensus paper warning that even though *highly capable AI*[13] may emerge within the next decade, current regulatory approaches fail to match the scale of this potential transformation. The paper also highlights critical gaps in AI safety research, noting that **only an estimated 1-3% of AI publications focus on safety**, and **no institutions currently exist with the authority or resources to prevent misuse or reckless development of powerful AI systems**. This disparity creates a clear opportunity for governments, multilateral organizations, and independent oversight bodies to fund AI safety research at a scale proportionate to the risks posed by advanced AI systems. The experts warn that AI could soon become highly capable in hacking, social manipulation, and strategic planning, escalating risks of large-scale cybercrime, autonomous weapons use, and even existential threats. Without meaningful intervention, AI systems could undermine trust in institutions, disrupt economies, and create unpredictable global security challenges.

The rising frequency of AI-related incidents provides **concrete evidence** that these risks are no longer science fiction. Rapid capability advancements in frontier models have made previously theoretical risks into practical possibilities. Real-world incidents involving AI systems provide shared reference points for discussing risks. According to the AI Incident Database, which tracks incidents related to the misuse of AI, 223 incidents were reported in 2024, a 49% increase over the 149 incidents that were reported in 2023 and over twice as many incidents as were reported in 2022. Since 2013, AI incidents have grown by over twentyfold. Similarly, the OECD's AI Incidents Monitors listed 5143 incidents and hazards in 2024, up from 373 in 2021.

Drawing attention to incidents reported in the AI Incident Database between October and November 2024 will help illustrate a broad range of issues, from errors in high-stakes systems, such as a transcription tool fabricating medical records[14] and inaccuracies introduced by ChatGPT in a child protection court report[15], to the misuse of AI in disinformation campaigns. For example, fabricated and AI-generated videos targeted Moldova's elections[16] and a US Vice-Presidential candidate.[17] Social media platforms like TikTok have hosted

---

with limitations. Voluntary commitments alone cannot address the full spectrum of risks posed by frontier AI systems, nor can they provide the enforceability needed to ensure consistent adherence across the industry.

[11] 2,778 researchers who had published in top-tier artificial intelligence (AI) venues gave predictions on the pace of AI progress and the nature and impacts of advanced AI systems.

[12] Namely: *Yoshua Bengio, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Trevor Darrell, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, Gillian Hadfield, Jeff Clune, Tegan Maharaj, Frank Hutter, Atılım Güneş Baydin, Sheila McIlraith, Qiqi Gao, Ashwin Acharya, David Krueger, Anca Dragan, Philip Torr, Stuart Russell, Daniel Kahneman, Jan Brauner , and Sören Mindermann*

[13] *highly capable AI* refers to *systems that surpass human abilities in critical domains*

[14] *Incident 827*

[15] *Incident 807*

[16] *Incident 840*

[17] *Incident 824*

AI-generated hate speech and propaganda[18], while China has reported waves of AI-driven fraud and disinformation[19]. These incidents reveal how AI can be used to manipulate narratives, threaten democratic processes, and erode public trust. The growing consensus around AI risks is shaped by real-world incidents, which provide a shared foundation for discussions on governance and safety. Personal harm caused by AI misuse is another key area of concern. For instance, students have used AI to generate explicit images of classmates[20], and explicit deepfake images have targeted underage students[21]. Additionally, facial recognition technology has resulted in wrongful arrests[22], and AI-generated misinformation has falsely implicated individuals in criminal activities[23]. Importantly, the consequences of these risks can be long-lasting and irrevocable[24].

Experts call for **governments to take a more proactive regulatory stance**. Actions could include establishing fast-acting AI oversight institutions with significantly larger budgets, mandating rigorous risk assessments with enforceable consequences, and requiring AI companies to demonstrate safety before deployment by adopting "safety cases" similar to those used in industries like aviation. For exceptionally powerful AI, the authors of a consensus paper propose that governments must be proactive by introducing licensing requirements, restricting AI autonomy in key societal decisions, enforcing strong security controls against potential misuse, and implementing automatic mitigation measures that activate when AI models reach certain capability milestones. A statement has also been signed by hundreds of executives and academics, including the chief executives of Google DeepMind, OpenAI, and Anthropic, all calling for stronger oversight [25].

3.3 Growing Public Awareness

In recent years, media outlets have increasingly focused on AI developments and their implications. This heightened coverage has led to a more informed public discourse about the need for oversight and has created political space for more ambitious governance initiatives. Civil society organizations have also become more sophisticated in their engagement with AI policy issues. Organizations focusing on AI ethics, safety, and governance have proliferated, contributing valuable perspectives to policy discussions. This broadened stakeholder involvement helps to ensure that emerging governance frameworks address the full spectrum of societal impacts.

The 2025 AI Action Summit in France built on growing public awareness and the increasing sophistication of global discussions on AI governance. In preparation for the Summit, two extensive consultation processes involving over 11,000 citizens and 200 experts were organized. These consultations revealed clear demands for inclusivity, fairness, and long-term governance strategies to shape AI for public good. Citizens and experts aligned on key goals—for example, harmonizing global AI governance, addressing societal challenges like

[18] *Incidents 809 and 810*
[19] *Incident 834*
[20] *Incident 848*
[21] *Incident 812*
[22] *Incidents 815 and 816*
[23] *Incidents 825 and 816*
[24] *As just one example, algorithmic housing discrimination and Incident 844*
[25] declaring "Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war."

labor market disruptions, leveraging AI for environmental sustainability, and ensuring equitable access to AI's benefits.

3.4 Increased Participation from Global Stakeholders

Over the past several years, countries around the world have become increasingly involved in AI governance efforts. Though high-level AI governance conversations have largely been dominated by the US, the UK, the EU, and China, there is a groundswell of AI-governance interest and activity from countries in the Global South. This increased global participation is crucial for creating inclusive AI governance frameworks and agreements that advance AI innovation and safety for the entire world.

Some recent highlights include the UN's High-level Advisory Body on Artificial Intelligence and its September 2024 final report that identified a pressing need to include more nations in AI future governance efforts. The Body itself comprised experts and senior government officials from 33 countries. Similarly, the UN's **Global Digital Compact** released in the same month emphasized AI's transformative potential and the need for a truly global partnership on AI. 143 UN member states voted for the Compact, amongst other initiatives, endorsing the need for global AI efforts to foster international development, ensure diverse languages and cultures are represented in AI systems, and build global technical capacity.

Other multilateral organizations in the Global South have also started to focus on AI development and strategy. The **African Union** released its **Continental Artificial Intelligence Strategy** in July 2024, for example, while Rwanda and Singapore recently published an **AI Playbook for Small States**. Organizers of the AI Action Summit also dedicated space for African nations and stakeholders to showcase regional AI initiatives and articulate continent-specific priorities for equitable AI development. If AI governance efforts are to be useful, impactful, and effective on a global scale, stakeholders must focus on the issues that are important and relevant for users around the world. These efforts to grow AI governance capacity beyond traditional power centers will thus help build consensus on priority topics regarding advanced AI systems within geographic regions (or similar international subgroups). With clarified priorities and centralized representation, these interest groups can better advocate for vibrant AI development in the larger coordination and governance mechanisms described in this review.

3.5 Potential for Safety Investment Growth

As institutional momentum and resources for making AI systems more powerful continue to surge, investments in AI safety, while comparatively modest, are receiving growing recognition for their importance.

In the United States, a consortium led by OpenAI, SoftBank, Oracle, and MGX has announced an unprecedented $500 billion commitment to fund **Stargate Project**, an initiative focused on expanding AI capabilities through advanced data centers and computational infrastructure. The European response to this competitive landscape is taking shape through multiple channels. The European Union established its €2 billion **AI Factories initiative** to enhance research and industrial applications, while France has separately unveiled plans for €109 billion in AI investments over the coming years. Meanwhile, China continues to strengthen its position in the global AI landscape, with an $8.2 billion **National AI Industry Investment Fund** to develop its domestic AI ecosystem. Chinese companies are making significant advances, with DeepSeek developing competitive AI models and Huawei investing substantially in chip technology to challenge Nvidia's market leadership.

These massive investments across major economic powers demonstrate a concerning trend: while vast resources are being allocated to expand AI capabilities, few resources are earmarked for AI safety. Still, there is clearly an appetite for increased funding, especially for critical AI-safety focused research: the number of academic papers regarding AI safety grew by over 300% from 2018-2023, with about 45,000 English-language articles about AI safety published during this timespan. Current gaps in funding for AI safety present a unique opportunity for supercharging public and private investments in the field.

Furthermore, safety techniques have proven to be not only conducive to responsible AI but also economically valuable. For instance, Reinforcement Learning from Human Feedback (RLHF), originally developed as a safety alignment technique, has become a mainstream approach for improving AI system performance and usability across the industry, demonstrating that safety research can yield significant downstream commercial benefits.

As a result, several nascent initiatives are directing funding towards safe AI and beneficial applications. The French government, in partnership with technology companies including Google and Salesforce and several philanthropic organizations, announced the creation of **Current AI**, a foundation dedicated to developing AI "public goods" with an initial endowment of $400 million and ambitions to raise $2.5 billion over five years. This multi-stakeholder initiative aims to support the development of public datasets, smaller-scale specialized AI models, transparent AI systems, and an open-source AI ecosystem, with participation from diverse nations including Germany, Finland, Switzerland, Chile, Kenya, and Nigeria. In the United Kingdom, the **Advanced Research and Invention Agency** (ARIA) has also emerged as a key player working to close this funding gap, directing resources toward high-risk, high-reward research including AI safety innovations that might otherwise remain underfunded.

4. Four Pillars for Effective Global AI Governance

The growing disparity between investments in AI capabilities and AI safety highlighted above has created an urgent need for structured, proactive governance efforts; this requires a coordinated, scientific approach that moves beyond voluntary principles to enforceable policies, ensuring that AI safety progresses in tandem with technological advancements. The following four pillars outline a comprehensive strategy to transform AI governance from a disjointed system into a scientifically rigorous, enforceable, and globally coordinated effort. These pillars address the core weaknesses in current governance structures by establishing a foundation for AI risk assessment, implementing regulatory frameworks, and strengthening institutional cooperation.

4.1 Pillar 1: Institutional Architecture

4.1.1 Formalizing the AI Summit Series

Background

International coordination on AI governance requires structured diplomatic platforms. The AI Summit Series, launched at Bletchley Park in 2023, continued in Seoul in 2024, and expanded through the Paris AI Action Summit in 2025, is the first high-level political process primarily focused on frontier AI safety. These summits achieved major milestones, including the first international political statements on AI safety, voluntary safety commitments from major developers, and the creation of national AI Safety Institutes in leading economies.

The Summit Series faces significant structural challenges as it matures. The Series' initial strengths, such as flexibility and informality, have become sources of fragility. As noted in Section 2.1, the proliferation of parallel initiatives and the absence of robust global coordination mechanisms risk fragmentation and diluted legitimacy. The shift from Bletchley's focused group of 28 safety-engaged countries to Paris's broader participation of over 90 states introduced new tensions between technical depth and political inclusivity, especially when the final statement failed to secure consensus from major players like the United States and the United Kingdom (see Section 2.2).

There is a pressing need to preserve AI safety as a core pillar of the AI Summit Series. In light of the withdrawal of safety language in the Paris Summit's final statement (see Section 2.2), alongside a broader pivot toward acceleration-focused agendas, maintaining focus on safety is critical. The evolving institutional landscape and the shifting posture of major governments add uncertainty to the long-term coherence of this process.

The AI Summit Series lacks essential governance mechanisms for long-term success. It currently operates without institutional continuity, accountability mechanisms, and a clear scope (Velasco et al., 2025). Without better-defined mandates and alignment with complementary forums, such as the International Network

of AI Safety Institutes, there is a risk of duplication or drift. As the only venue combining high-level political attention with a focus on frontier risks, the Summit Series must now evolve into a more structured, transparent, and enduring diplomatic process.

Recommendations:

1. **Turn the AI Summit Series into a sustained, structured, and inclusive international process**. Establish a focused, advanced-AI track that brings together countries with advanced AI capabilities and major developers to coordinate safety governance, and a broader public-interest track that addresses the societal impacts of AI. Each track should produce concrete deliverables rather than general declarations. To ensure global legitimacy and effectiveness, both tracks should be linked through formal coordination mechanisms and inclusive feedback channels. Both tracks should meet at high-level summits once per year, supported by technical-ministerial meetings every six months. The annual summits would provide a predictable forum for high-level political engagement, strategic decision-making, and major announcements. Interim meetings would allow participants to track developments, review commitment implementation, and prepare substantive input for the main summits.

2. **Create an institutional framework by the next AI Summit in India.** Establish a **rotating secretariat** with dedicated staff to provide logistical support, track commitments, coordinate stakeholders, and ensure agenda continuity. The secretariat would serve as the central coordination node, preserving institutional memory and facilitating alignment between Summits led by different host countries. It would also support consistency in reporting, follow-up, and working group facilitation, ensuring that each Summit builds on the commitments and declarations of the previous one. Hosts should be selected at least two years in advance through a geographically balanced rotation mechanism, with criteria for continuity and preparation capacity clearly defined. Form a **permanent steering committee** with representatives from AI safety institutes, regulatory agencies, international organizations, civil society organizations, and academia to align Summit priorities with other governance processes such as the UN, INASI, G7/G20, and regional AI initiatives. The steering committee should identify high-priority areas for advancement across summits, including safety evaluation protocols, incident response infrastructure and alignment with emerging international standards and red lines.

3. **Consolidate follow-up and accountability mechanisms under the rotating secretariat and steering committee.** The secretariat should oversee a standardized process for tracking national and institutional commitments, requiring participants to submit progress updates 45 days prior to each summit. These inputs should inform a publicly available post-summit report within 60 days that details implementation progress, identifies barriers, and outlines next steps. This process should reinforce transparency, enable civil society oversight, and ensure that working group outputs and summit deliverables are meaningfully integrated into future agendas.

4. **Establish four permanent working groups with clearly defined mandates.** Each group should consist of independent experts, civil society organizations, private sector representatives, and national delegates. These groups should operate year-round and present progress updates at each summit. Working groups should cover:

a. **WG1 on Coordinating Global Standards**, in charge of mapping existing and emerging legal, regulatory, and voluntary instruments, identifying gaps and overlaps, and developing mechanisms to resolve conflicts between potentially divergent approaches.
b. **WG2 on Risk Thresholds and Mitigation Measures**, in charge of defining and presenting capability-based thresholds that trigger escalating regulatory scrutiny or restrictions (especially based on triggers such as autonomy or dual-use potential), and presenting technical and organizational mitigation measures (see also 4.3.2).
c. **WG3 on Verification Mechanisms**, responsible for developing protocols and tools to independently verify claims about AI systems across their lifecycle. This could include developing cross-border inspection procedures, cryptographic verification methods, and standardized reporting mechanisms to build mutual trust between stakeholders.
d. **WG4 on Global Access and Benefit Sharing**, in charge of developing concrete frameworks for equitable distribution of AI benefits, including profit-sharing, access to compute, training data, model capabilities, and technical expertise. They would help establish metrics for redistribution and alignment with international development goals.

These working groups should coordinate closely with the scientific assessment mechanisms outlined in Section 4.2.1 and Pillar III (Scientific Understanding) to ensure alignment between technical analysis and governance actions.

4.1.2 Strengthening the Network of AI Safety Institutes

Background

National AI Safety Institutes (AISIs) play a central role in frontier model risk assessment. While summits and political bodies articulate shared principles and high-level commitments, AISIs work at the technical level to evaluate frontier systems and are well-positioned to translate safety insights into actionable

policy inputs. Since their launch, AISIs have begun to form bilateral and multilateral partnerships, evaluating risks, developing methodologies, and informing governance. The launch of the International Network of AI Safety Institutes (INASI) after the Seoul Summit (2024) formalized global coordination among these bodies.

International coordination of AISIs is essential to prevent fragmentation and regulatory arbitrage. Global AI development remains concentrated in a few countries, and not all nations can independently build robust evaluation capacity. Isolated national efforts risk inefficiency and divergence in safety standards. A coordinated network of AISIs offers a promising solution to align technical evaluations across jurisdictions while supporting under-resourced regions. However, this ambition requires robust information-sharing mechanisms, common methodologies, and an inclusive institutional structure. Without these, the effectiveness and legitimacy of AISIs will remain limited.

Political momentum for enhanced coordination is growing but requires institutional support. The AI Seoul Summit and the G7 Hiroshima Process have demonstrated a baseline level of commitment from frontier AI states. Yet, in anticipation of accelerated AI progress and potential crunch scenarios, further institutional development is needed to ensure a well-functioning safety infrastructure for when it becomes most critical.

Recommendations:

1. **AI Safety Institutes should work toward the development of international AI safety standards by 2027.** In particular, AISIs should collaborate to create best practices in AI evaluations risk thresholds. These efforts, carried out in tandem with international standard-setting bodies and other organizations involved in standards development, should include standardized testing protocols for all models exceeding defined capability threshold, risk assessment, and safety engineering practices. Following joint testing exercises and collaborative technical projects to build shared evaluation suites for AI models, systems, and agents, states should also foster the development of guidelines for third-party model evaluations. A coordinated evaluation program should be launched by mid-2026, supported by a shared repository of results with appropriate security and access controls.

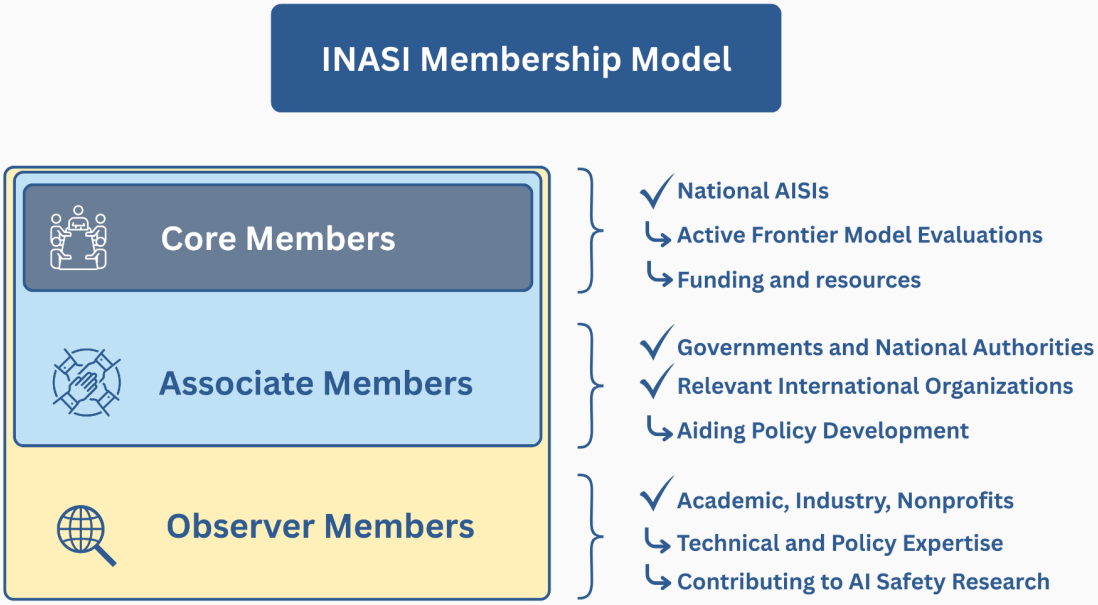
2. **Create a structured membership model for INASI with clearly defined responsibilities.** Define three categories:
    a. **Core members —** National AISIs with active frontier model evaluations. Responsible for driving network strategy and overseeing its activities, chairing working groups, and providing resources.
    b. **Associate members —** Governments without AISI equivalents, national authorities beyond AISI whose part of the mandate may be relevant to AI governance, relevant offices of international organizations working on AI governance. Responsible for sharing relevant policy developments and facilitating information-exchange, ensuring domestic uptake and awareness of INASI priorities and work.
    c. **Observer members —** Academic institutions, research labs, relevant divisions of companies from the AI ecosystem, or nonprofits contributing to safety research. Responsible for providing technical and policy expertise and input, participating in designated working groups, participating in or leading INASI-sponsored projects.

    This structure enables INASI to remain open and inclusive while ensuring that decision-making and technical responsibilities rest with institutions best positioned to fulfill them. Broader participation is

made possible without compromising evaluation integrity or overburdening limited resources. A shared charter should be established to outline baseline governance structures, expectations, and opportunities for specialization. Institutes can then complement one another's strengths, for example, with some focusing on agentic testing, while others take the lead on red-teaming or bio-risk evaluations to maximize global expertise and support capability development.

3. **Establish regional AISI hubs by 2027**. These hubs, covering all 6 regions as defined by the United Nations system, should focus on capacity building, training, and regional research coordination, and train at least 100 technical specialists annually through standardized yet locally adapted curricula. Regional AISIs should actively contribute to international technical processes, engage with local standards bodies, and secure representation in global governance forums to ensure international harmonization. For countries without national AISIs, these regional hubs can provide access to shared infrastructure and expertise, ensuring meaningful participation in safety efforts.

4. **Establish an international AI risk information-sharing system**. By the end of 2026, develop secure infrastructure including encrypted communication channels, standardized reporting templates, and legal frameworks to support the confidential exchange of sensitive information about model capabilities, risks, and mitigations. This foundation should enable systematic information-sharing between AISIs for critical vulnerabilities, with cross-jurisdictional notification of safety breaches. Building on this infrastructure, create **specialized Information Sharing and Analysis Centers (ISACs) to serve as trusted operational entities** fulfilling several functions: a Risk Communication Function for immediate vulnerability sharing, a Knowledge Management Function maintaining secure repositories of evaluation results, and an Analysis and Early Warning Function for horizon scanning. These ISACs should be integrated within the International Network of AISIs, providing standardized protocols for collecting, anonymizing, and sharing advanced AI risk information while facilitating cross-border coordination in response to emerging threats. The centers should prioritize data privacy, cybersecurity, and proprietary information protection while enabling timely access to risk-relevant information for policymakers, researchers, and evaluators. Where able to, ISACs should be encouraged to share relevant information about capabilities and risks with the proposed UN independent scientific panel (see section 4.2.1).



4.1.3 The Role of the United Nations

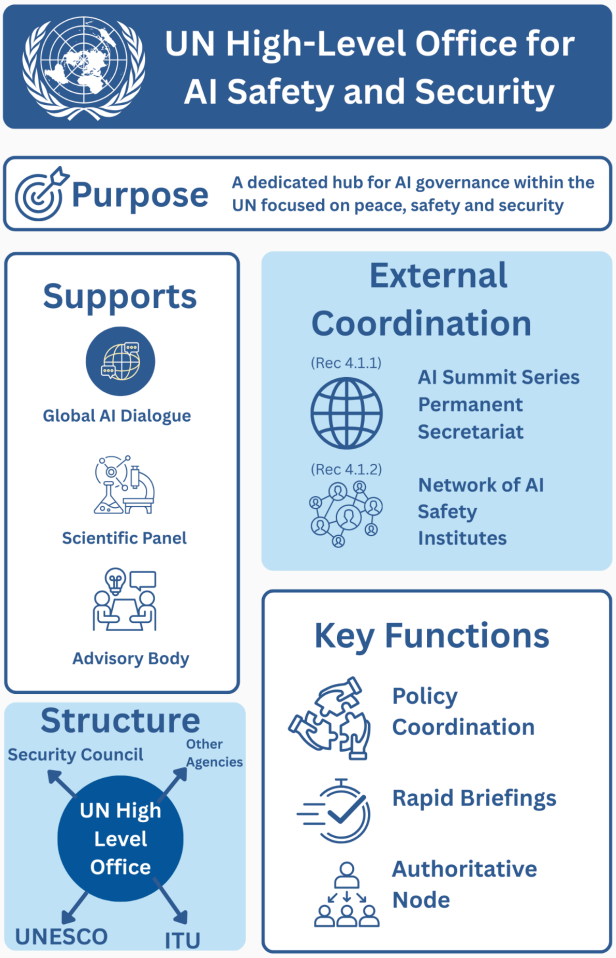
Background

The United Nations provides immense potential—but also raises challenges—for global AI governance and coordination. As the preeminent multilateral forum with universal membership, the UN plays a crucial role in establishing legitimate global governance frameworks for emerging technologies. Despite its unmatched convening power and legitimacy, however, the UN's complex bureaucratic structure and consensus-based decision-making present challenges for developing nimble responses to rapidly evolving AI capabilities.

Recent UN initiatives have formally acknowledged the risks of advanced AI systems. The UN has responded to growing international momentum on AI governance through several initiatives, including the March 2024 resolution on "Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development," adopted by consensus. The resolution formally acknowledged the potential large-scale risks posed by advanced AI systems and affirmed the responsibility of all states to ensure proper oversight and regulation. Building on this foundation, the Summit for the Future in September 2024 presented the Digital Global Compact, which established two new AI-focused initiatives: the Global Dialogue on AI, an international forum where countries and stakeholders collaborate to foster shared approaches to AI governance, and the Independent International Scientific Panel on AI, a group of experts providing assessments of AI developments to inform global decision-making.

Recommendations:

1. **Establish a dedicated UN High-Level Office for AI Safety & Security**. The existing UN Office for Digital and Emerging Technologies (ODET) plays an important role in supporting international cooperation on digital and emerging technologies, including AI. However, AI is such a transformative and impactful technology, with such distinct safety and security challenges, that it merits its own High-Level Office. This division would allow ODET to focus on a range of other digital and emerging technology issues. This High-Level Office would provide an institutional home, stable funding, and knowledge continuity for the UN Global Dialogue on AI, the Independent Scientific Panel on AI, and the High-Level Advisory Panel. The Office should be adequately staffed and report directly to the Secretary-General. It should coordinate policy inputs horizontally across UNESCO, ITU, and other specialised agencies; channel their findings vertically to Member-State focal points; and, when necessary, fast-track briefings to the Security Council. Some competencies of ODET related to AI risks should thus be transferred to this new High-Level Office, allowing ODET to continue its broader work on technology access, standards, and socio-economic impacts. By acting as the single authoritative node for AI safety and security policy inside the UN, and linking to external processes such as the AI Summit Series secretariat (Rec 4.1.1) and the Network of AI Safety Institutes (Rec 4.1.2) the Office would ensure that the UN system's convening power is focused on the peace-and-security and safe development of advanced AI.



2. **Structure a comprehensive global AI capacity-building strategy** through two complementary approaches: 1) establish a dedicated track within the UN Global Dialogue on AI to clarify the needs, objectives, and principles for building technical and administrative capacity to govern AI, and 2) implement a pilot program jointly led by the UN Development Programme and the International Telecommunication Union to develop administrative capacity in priority countries by 2027 (for technical capacity-building elements, see section 4.4.1). The program should train at least 200 government officials annually, establish rapid-response technical assistance teams, and

collaborate closely with AISIs and regional AISI hubs (as outlined in Recommendation 4.1.2).

3. **Negotiate a binding international framework convention for AI safety by 2027.** Undertake a comprehensive legal analysis to clarify the applicability of existing international law to advanced AI systems, followed by formal negotiations leading to a binding international framework convention. Drawing from the climate model, where a framework convention (the UNFCCC established core principles that were later enhanced through additional protocols (e.g. the 1992 Kyoto Protocol), a framework convention for AI should outline the foundational objectives, processes, and principles of international governance for tackling the safety, security, and economic issues associated with advanced AI. Subsequent protocols should establish clear state obligations regarding frontier AI development and risk management, align with the risk classification and threshold frameworks proposed in Recommendation 4.3.2, and define enforceable measures for preventing catastrophic AI misuse.

4.2 Pillar II: Scientific Understanding

4.2.1 Increasing Scientific Assessment of Advanced AI

Background

Shared scientific understanding forms the backbone of international governance. In climate policy, the Intergovernmental Panel on Climate Change has significantly influenced how nations conceptualize and address climate challenges. In the AI domain, an international scientific panel could foster consensus on the risks and trajectories of general-purpose AI systems, helping shape coordinated global policy responses.

The first International AI Safety Report marked a milestone, but requires institutionalization. This vision gained traction with the January 2025 release of the first report, authored by an expert group chaired by Yoshua Bengio and temporarily hosted by the UK government. The report provided a comprehensive assessment of advanced AI risks, ranging from deepfakes to misuse in chemical and biological contexts. However, for scientific risk analysis to play a sustained role in AI governance, it must evolve into a regular and institutionalized process. Efforts are underway to anchor this process in a more permanent international framework.

Budding efforts at the United Nations to support scientific assessment of advanced AI are promising. The UN's draft proposals for an Independent International Scientific Panel on AI envision a scientific and technical board tasked with publishing regular reports about opportunities and risks from AI, including evidence-based capability and risk assessments. The Panel would feature an Executive Committee composed of twenty global experts appointed by the Secretary-General in addition to a larger Advisory Committee composed of forty members elected by the General Assembly with consideration for geographical, gender, and disciplinary balance. These efforts represent a large step in the right direction to ensure continued and internationally-supported scientific investigation of risks from advanced AI.

Significant challenges persist in effective international scientific consensus-building of AI. Traditional peer review is too slow for fast-moving research, and political discussions can trail behind technical developments. Meanwhile, critical findings now often first appear in preprints, technical blogs, and industry reports. It is essential that any long-term mechanism retain editorial independence, integrate insights from those diverse research sources, and establish the agility needed to keep pace with rapidly evolving AI capabilities, all so that policymakers have timely access to the latest evidence. An increasing imbalance between well-resourced private-sector research labs and underfunded public research institutions exacerbates transparency and access gaps, concentrating frontier model capabilities and insights within a handful of corporate actors. This concentration heightens risks and limits the inclusivity of global safety assessments.

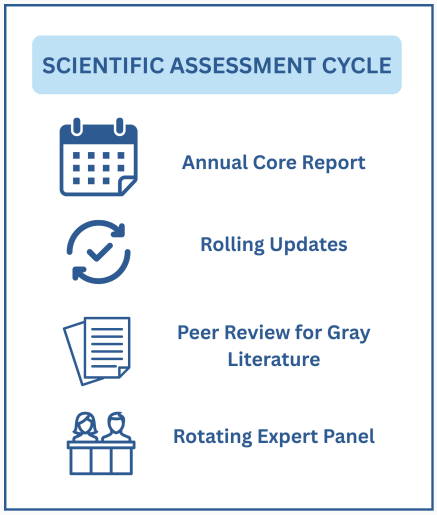
Recommendations:

1. **Institutionalize the International AI Safety Report process within the UN Independent International Scientific Panel on AI.** The Scientific Panel should strike a balance between the political legitimacy enabled by UN and member state support and the scientific legitimacy created

by a robust and independent process for producing evidence. This process would be an impactful, transparent, and globally representative scientific endeavor if executed well; international policymakers should thus prioritize the panel's independence from political considerations. It should include mechanisms to rapidly synthesize emerging evidence from non-traditional sources such as preprints, technical blogs, and corporate disclosures, and translate it into a comprehensive report on the state of the art of AI Safety.

2. **Create Dedicated Working Groups for Advanced AI Safety and Security.** Establish one or more cross-disciplinary WGs under the Expert Committee dedicated to building shared scientific understanding on risks from advanced AI. These specialized WGs should be responsible for authoring the safety and security sections of broader panel-wide reports. Due to the rapid pace of AI development, these sections could be updated more frequently than other report components to ensure timely assessments. The WGs should also have the ability to issue concise technical bulletins in response to novel developments, providing rapid expert analysis when needed.

3. **Implement Rolling and Ad hoc Publication Cycles.** Deliver one comprehensive report each year, covering the full AI landscape and policy implications, plus one mid-year thematic update focusing on emerging high-risk areas. The Expert Committee should also be authorized through specific protocols to request the drafting of technical bulletins after trigger events such as significant scientific breakthroughs in AI capabilities or safety research, newly documented emergent behaviors in advanced AI systems, substantial shifts in technical consensus about risk levels, unexpected acceleration in model scaling or deployment, or the discovery of previously unidentified failure modes. Approved bulletins must be published in open-access format within 7 days, translated into all UN official languages. Summaries tailored for policymakers, technical audiences, and civil society should be distributed through targeted webinars, infographics, and policy briefs.



4. **Establish Formal Communication Channels with Key Stakeholders.** Appoint liaison officers within the Panel Secretariat to serve as points of contact with AI Safety Institutes (especially through our proposed AI Safety Information Sharing and Analysis Centers), leading AI research laboratories and companies, academic research centers, and industry consortia. These relationships will ensure comprehensive data collection for panel reports. Liaison officers will manage secure communication channels, handle sensitive data under UN protocols, and coordinate regular information exchanges. Outputs from these engagements should feed directly into the Panel's working group agendas and reporting cycle.

4.2.2 Boosting Research Funding and Coordination

Background

AI safety research faces severe underinvestment and fragmentation. Investment in AI safety research is essential for developing solutions to technical problems and risk mitigation strategies before the deployment of increasingly capable systems. Currently, safety research constitutes only approximately 1-3% of overall AI research publications. In terms of funding, data is severely lacking, but most estimates put AI safety investment at no more than hundreds of millions of dollars per year compared with pledges of hundreds of billions for overall AI capability advancement. This imbalance has developed as commercial and governmental interests have prioritized capability advancement over safety considerations, creating a widening disparity between technological advancement and risk mitigation.

International scientific collaboration models offer promising templates for AI safety research. Specialized fields have demonstrated transformative potential through initiatives like CERN and Horizon Europe, providing models for how coordinated research networks might accelerate progress on AI safety challenges. Key stakeholders in setting up such efforts include academic researchers, independent research organizations, industry research labs, government funding agencies, and philanthropic foundations supporting safety work.

The double-edged nature of AI safety findings creates potential tensions between collaboration and security concerns. Experts have called attention to the blurred boundary between safety and capabilities research, emphasizing the need for practical methods to distinguish research directions that reduce risk from those that inadvertently accelerate dangerous capabilities. Research shows that interpretability techniques, for instance, help understand model internals for safety purposes but can also reveal pathways to enhance dangerous capabilities. Building on such proposals, new international coordination efforts should seek to operationalize this distinction and align funding mechanisms accordingly. While examples of successful cross-border collaboration exist, the potential harmful applications of AI safety research complicate international coordination efforts. This dual potential creates a potential barrier to collaboration, since states might be hesitant to participate in collaborative research projects that could equalize technological disparities they view as advantageous.

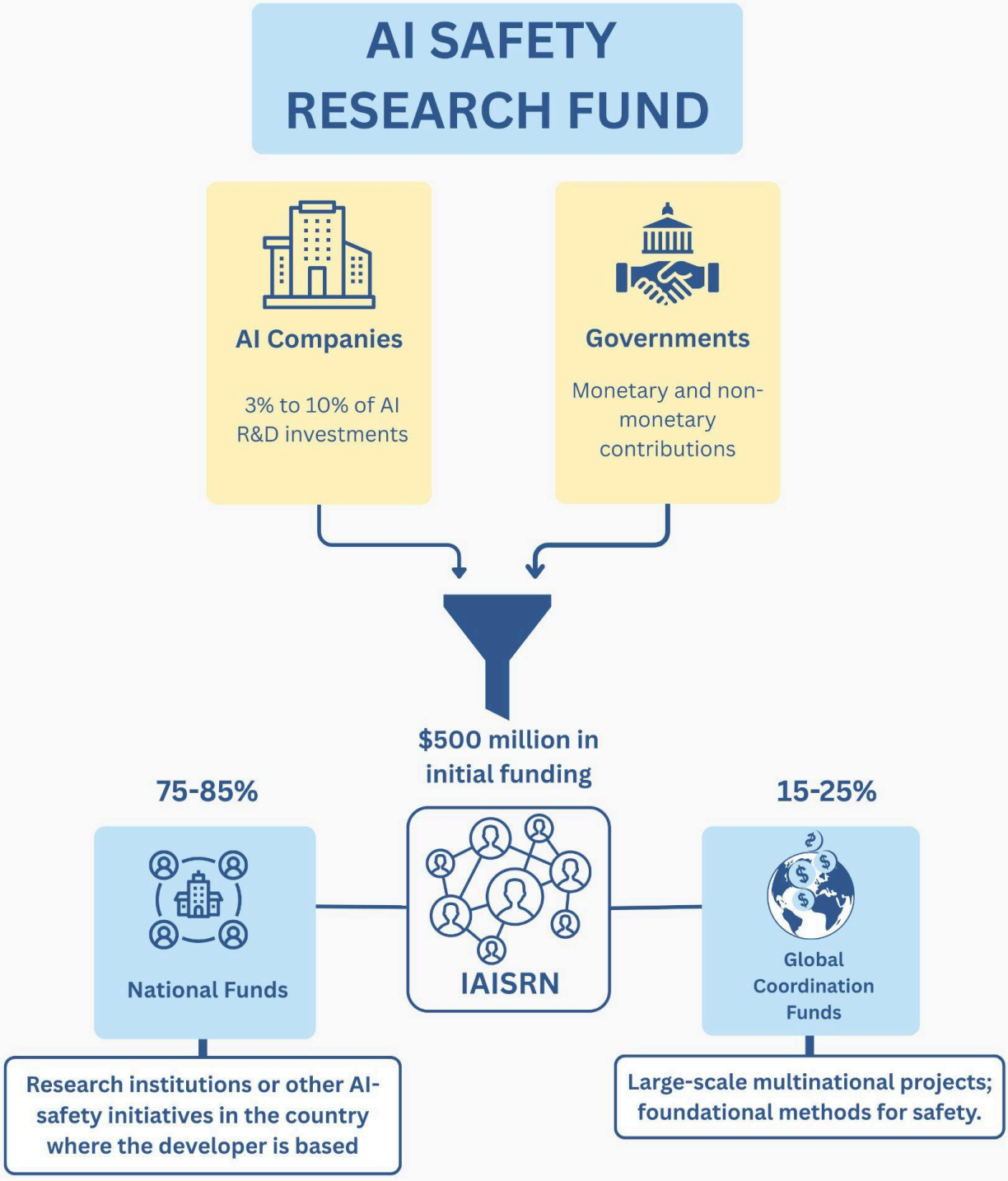
Recommendations:

1. **Establish an International AI Safety Research Network (IAISRN) by 2026, supported by $500 million in initial funding** (as explained further below). Structured as a non-governmental or quasi-governmental organization, this Network would coordinate independent scientific research on the risks of advanced AI systems. It would be responsible for developing and maintaining an integrated research agenda for AI safety and security and launching large-scale multinational projects. This agenda should include clearly defined milestones and metrics, a rolling 1-2 year roadmap of technical challenges, and mechanisms to adjust priorities in response to emerging capabilities. The agenda should be updated regularly and presented at key global forums such as the AI Summits. As a foundational project, the IAISRN should launch the **AI Safety-Capability Differentiation Project** to distinguish between safety research streams that inadvertently advance capabilities and those that enhance safety without proliferation risks,

producing a taxonomy and an evaluation framework. It would also offer practical guidance to researchers and institutions on prioritizing capability-neutral safety pathways.

2. **Require qualified AI companies to contribute to AI safety research through a structured funding mechanism.** Companies developing models above specified compute levels (e.g., systems trained using >10^25 FLOP) would be required to make a contribution to an AI Safety Research Fund of 3% to 10% of the company's AI R&D budget—a substantial amount yet proportional to sector R&D expenditures. Governments could also establish frameworks allowing industry contributions in non-monetary resources such as compute access, data, algorithms, or researcher time when these alternatives deliver equivalent or greater value. 75-85% of the collected funds would be allocated to research institutions, projects, or other AI-safety initiatives in the country where the developer is based—though the exact disbursement or allocation process requires further clarification. This allocation would provide a strong incentive for states to approve a measure requiring such substantial investments from industry. The remaining 15-25% of proceeds would be earmarked to a global fund for large-scale multinational projects, such as those led by the International AI Safety Research Network. This global fund would be governed by equal voting rights among participating states alongside a scientific advisory group for technical recommendations. In addition to AI companies, governments wishing to demonstrate their dedication to AI safety should be encouraged to contribute to the global fund voluntarily. This structure aims to balance the need for globally pooled investments in high-impact multinational projects with incentives for national investment in AI safety infrastructure.

3. **Establish regional centers of excellence for AI safety research**. Connect existing centers of expertise (such as the GPAI Expert Support Centres) to the International AI Safety Research Network, and launch new centers to ensure comprehensive coverage of critical AI safety research areas. Each center should specialize in a key domain such as mechanistic interpretability, alignment, robustness, etc. These centers should serve as technical anchors for the International Network's global research agenda, conduct cutting-edge safety research, and support regional talent development through fellowships, training, and exchange programs. Funding should be drawn from the AI-safety-focused research funds proposed above, with contributions from both industry and governments. The global fund should allocate resources to support the launch, operation, and sustained impact of these specialized hubs.



Centre pour la Sécurité de l'IA, 2025

4.3 Pillar III: Rules of the Road

4.3.1 Monitoring Corporate Commitments

Background

Current AI governance relies heavily on voluntary industry commitments. In recent years, a growing focus on accountability for frontier AI labs has led to increasingly specific commitments. There has been movement from early ethics principles to more specific commitments about risk assessment, red-teaming, and deployment safeguards, reflecting a growing recognition of concrete safety challenges. Still, the current landscape relies on voluntary commitments and self-reporting mechanisms from frontier AI developers, including pledges made at the Seoul Summit and through the G7 Hiroshima Process Code of Conduct. These voluntary commitments serve as crucial interim measures while binding international standards are being developed[26]. The monitoring of these commitments should align with and inform the work of AI Safety Institutes, the proposed International AI Safety Research Network (IAISRN), and the development of global safety standards.

Significant governance gaps undermine trust in advanced AI systems. While voluntary approaches have demonstrated some industry willingness to engage with safety concerns, the absence of standardized reporting requirements, independent verification mechanisms, and consequences for non-compliance creates significant oversight weaknesses. As just one example, several companies failed to deliver on their May 2024 commitment to publish frontier safety frameworks before the February 2025 Paris summit.

Certification and evaluation methodologies for AI remain underdeveloped across regions. Robust certification frameworks are an essential tool for verifying whether companies are fulfilling their obligations, whether through voluntary commitments, industry-led standards, or regulatory requirements. But existing conformity assessment frameworks are not yet adapted to the unique challenges of advanced AI. The methodologies used for certifying complex software systems must evolve to account for non-traditional evidence sources, emergent behaviors, and "black-box" characteristics of advanced AI. Current systems lack standardization around how competency, transparency, and assurance are evaluated. Without updated and internationally aligned accreditation structures, attempts to scale up AI evaluations risk fragmentation, duplication, or superficial assurance.

Recommendations:

1. **Develop an international system for AI companies' public reporting on risk management and responsible AI**, supported by advanced AI States to ensure global adoption. A standardized reporting template for AI companies to demonstrate commitment adherence should be developed, including for relevant requirements in the Seoul Frontier Safety Commitments, the G7 Hiroshima Process Code of Conduct, the EU's GPAI Code of Practice, and China's AI Safety Commitments. By using this template, organizations will be able to clarify which frameworks they're adhering to and how they're implementing them. Such a system would facilitate sharing of best practices across the industry, establish reputational incentives for responsible

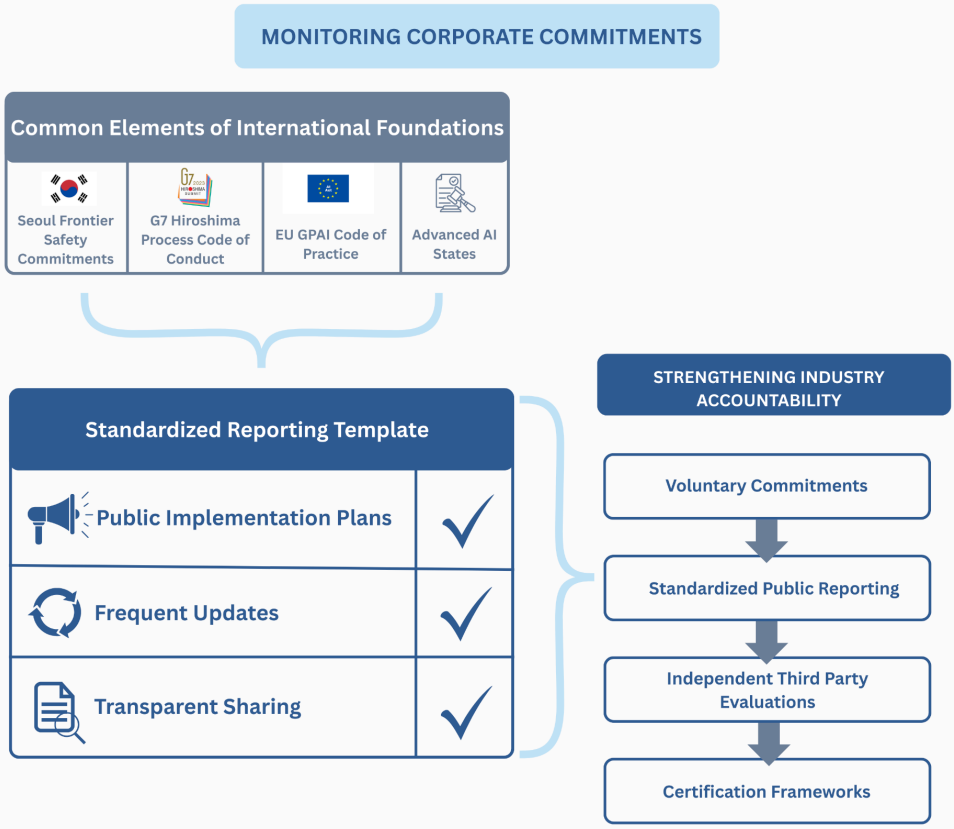
[26] See Section 4.1.2 on AISI coordination and Section 4.3.2 on red lines for unacceptable risks

development, and enable more effective identification of emerging risks and governance gaps.

2. **Build an international ecosystem of qualified independent third-party AI evaluators.** More independent AI evaluators are needed, and these evaluators need better access to the inner workings and technical details of AI models. On the first point, more resources must be allocated to grow the ecosystem of trusted third-party evaluators, as current nonprofit evaluation institutes like METR and Apollo Research are few in number, while AI Safety Institutes might not have the capacity or incentives to carry out comprehensive evaluations of all potentially risky frontier models. Some of this funding could be allocated from the proposal laid out in Section 4.2.2 above. On the second point, states could encourage AI developers through voluntary and legal measures to enable appropriate access for technical audits. As the ecosystem of evaluators develops and matures, states should collaborate to establish robust qualification criteria defining the technical expertise, independence, and other requirements for what constitutes a 'qualified' evaluator. These shared standards could form the foundation for international registries of qualified evaluators, enabling more efficient and trustworthy cross-border recognition.

3. **Promote the development of safety certification frameworks for high-risk AI systems**, including general-purpose models exceeding defined capability thresholds. These should incorporate dynamic monitoring systems that enable continuous oversight, particularly after significant model updates. Unlike traditional safety-critical sectors, the adaptive nature of AI models requires ongoing compliance mechanisms that evolve alongside system behavior. Requirements for certification should be regularly renewed to take into account emerging risks and responsible development and deployment practices.

4. **Support the establishment of an independent, NGO-led international Corporate AI Safety & Security Observatory by 2026 to evaluate AI companies' progress on their voluntary commitments and risk management practices.** The Observatory will (a) collect and analyze company reports, evaluation results, certification data, and incident reports from national regulators, accredited conformity-assessment bodies, and feeds such as the OECD AI-Incident Framework and NIST ARIA; and (b) publish quarterly Progress Reports to assess industry-wide progress, thereby maintaining consistent pressure to incentivize compliance and highlight areas needing improvement; and an Annual State of AI Incidents Report with sector-specific risk dashboards and early-warning indicators. Independence should be ensured through diversified five-year funding baseline public grants, a fixed 10 % share of the AI Safety Research Fund held in a trust insulated from political or commercial influence. With about 30 technical staff and secure data-sharing MoUs in place by 2026, the first Reports should be released no later than 2027. The Observatory should work and share its findings with the International Network of AISIs to avoid duplication of efforts.



4.3.2 Developing Red Lines for Unacceptable Risks

Background

Risk thresholds serve as predetermined risk levels that, when exceeded, trigger specific responses (e.g. to deploy more careful monitoring). "Red lines," on the other hand, represent specific capabilities or applications whose risks are so unacceptable that they should not be developed and/or deployed. Without clear thresholds and enforceable red lines, developers may enable dangerous capabilities, whether through deployment of deceptive agents, autonomous bio-threat design, or other catastrophic misuse. The implementation of such protective measures presents significant challenges. While countries may implement red lines through domestic legislation, the global nature of AI development means that companies could simply relocate development activities to jurisdictions with less stringent regulations. International agreements that define unacceptable AI behaviors and establish common guardrails across borders will be necessary to prevent companies from evading red lines through jurisdiction hopping.

Red lines enable concrete, proactive risk management without requiring risk-timeline consensus. By focusing discussion on specific dangerous capabilities and their corresponding detection methods and mitigations, red lines are a practical compromise between those who view advanced AI risks as imminent and those who consider them distant, allowing stakeholders to agree on necessary actions without consensus on when capabilities will emerge. By establishing these capability-response pairings in advance, safety measures

can be developed incrementally. This enables learning from implementation experiences and ensures that protective measures will be ready when needed rather than developed reactively after risks materialize.

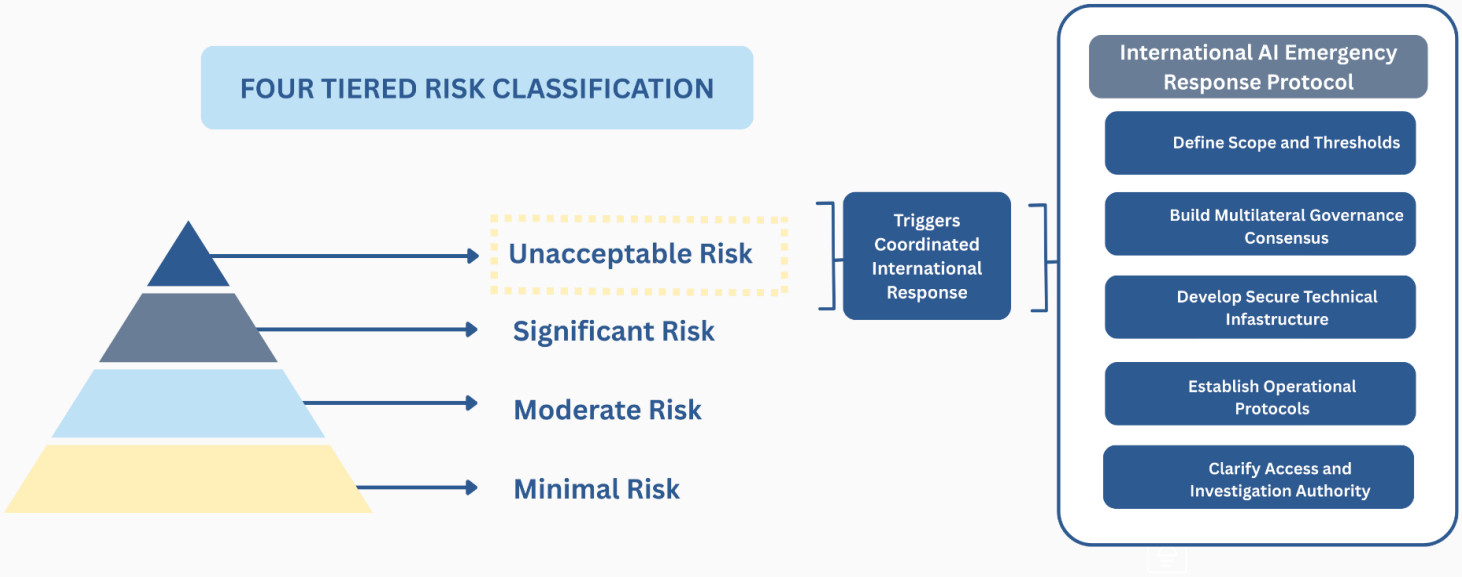
Tiered risk classifications have emerged across multiple regulatory frameworks but lack coordination. These frameworks, a critical part of efforts to define red lines, have gradually evolved from general ethical principles toward more specific technical criteria for capability assessment and risk categorization. Tiered risk classifications appear in the EU AI Act and voluntary industry standards, though with significant variation in definitions and enforcement mechanisms. While they have gained broad conceptual acceptance, the lack of common standards potentially enables regulatory arbitrage, allowing developers to relocate to jurisdictions with weaker oversight.

Recommendations:

1. **Establish international consensus for categories of unacceptable risks (or "red lines") to prevent large-scale harm.** States should develop international consensus on specific AI capabilities that should be subject to strict limitations, such as autonomous cyberattacks or self-replication. This categorization should be regularly reviewed and updated to account for emerging capabilities and potential threats. These red lines could be documented in political declarations and implemented through domestic authorities, building upon efforts like the Hiroshima Process Code of Conduct, the Bletchley Declaration, and the International Dialogue on AI Safety in Beijing.

2. **Reach consensus on a four-tier AI risk classification framework**. States should establish an international consensus for categorizing AI systems based on their risk level and capabilities. This classification would distinguish between **Minimal Risk** (minimal or no oversight), **Moderate Risk** (basic transparency), **Significant Risk** (comprehensive obligations), and **Unacceptable Risk** (prohibited development, i.e. red lines). Agreement on the content and application of these categories is critical to avoid regulatory divergence and facilitate coordinated oversight. Such alignment could be advanced through joint declarations at upcoming summits and would provide a foundation for coordinated international responses to AI development.

3. **Develop a binding international emergency response protocol for advanced AI systems by 2026.** This protocol should outline the procedures to be followed when defined AI risk thresholds or red lines are breached. To ensure coordinated and timely responses to serious AI incidents, the protocol should articulate early warning criteria and triggers for escalating alert levels based on technical indicators and threat severity as proxy for red lines capabilities. It should establish encrypted, real-time communication systems for use by authorized national agencies, AI Safety Institutes, and designated emergency coordination centers for rapid exchange of information about threat intelligence, incident disclosures, and containment measures (see also section 4.1.2). The protocol must define pre-authorized containment measures that can be automatically implemented at higher alert levels within a 12-hour window. Procedures for identifying which governments, safety institutes, and evaluators have temporary access to affected AI models for investigative and containment purposes should be specified, with strict non-proliferation protocols to prevent misuse. Finally, AI developers must be required to notify relevant government authorities within hours of identifying a severe risk threshold breach, including technical evidence, a timeline of events, and immediate actions taken.



4.4 Pillar IV: Ensuring Access

4.4.1 Sharing AI Benefits

Background

Equitable access to AI technologies is essential. Currently, AI infrastructure, development, and deployment remain largely concentrated in a select few countries, primarily the United States and China. This distribution enables significant disparities in access to computing resources, talent networks, and economic opportunities. Without deliberate mechanisms for benefit sharing, AI could exacerbate existing inequalities rather than reduce them.

AI's labor market impacts create additional urgency for financial benefit-sharing programs. AI's potential to reallocate and replace human labor provides states extra incentive to quickly design and have ready for implementation financial benefit-sharing programs. For example, policymakers could attach windfall financial clauses to future data-center or energy-plant subsidies, with stipulations that such windfall proceeds would be shared nationally and internationally.

Diverse stakeholders have varying perspectives on AI benefit sharing and opposing interests. Actors have a wide range of perspectives and interests regarding the development and benefit-sharing of advanced AI systems. For example, nations controlling critical infrastructure, international organizations facilitating cooperation, developing countries seeking meaningful participation, and private companies with proprietary AI systems have very different views about what equitable access to AI benefits entails.

Initial efforts to address AI access imbalances remain insufficient. These include the United States' Partnership for Global Inclusivity on AI launched under President Biden in September 2024, China's Global AI

Initiative, and the UN's Global Digital Compact. These programs have started to outline frameworks for compute access, capacity building, and data assistance for developing nations, but these efforts appear distinctly insufficient to ensure that AI benefits are fairly distributed—for example, the US PGIAI was reliant on millions of dollars of uncertain federal funding, when larger efforts (in the billions of dollars) are likely necessary for widespread benefit sharing.

At the same time, some forms of global resource-sharing must be carefully managed to minimize proliferation risks. Projects involving model access or dual-use capabilities should be narrowly scoped to advance public-interest safety objectives while avoiding the dissemination of sensitive capabilities. These projects should rely on secure, privacy-preserving evaluation infrastructure to maintain trust and protect proprietary or high-risk data.

Recommendations:

1. **Establish a global financial AI benefits redistribution mechanism with clear allocation formulas, to be activated once AI generates a substantial share of global economic output.**. Create a framework requiring companies exceeding specific revenue thresholds from AI to contribute to a global fund under UN auspices, building on existing multilateral frameworks. For example, this might take shape as a fee on frontier AI revenue, with progressive rates from 1-3% of AI-derived profits, scaling up as AI's economic impact grows. This windfall could then allocate resources to universal basic income programs, AI education programs, and public-interest AI applications (see also O'Keefe et al., 2020).

2. **Create a Global AI Solutions Fund to incentivize AI development for developing economy challenges through a combination of advanced market commitments, innovation prizes, and matching funds**. The fund would operate through multi-tiered incentives including challenges for transformative solutions, smaller prizes for targeted technical achievements, and implementation awards for successful deployment. Advanced market commitments would guarantee purchase of successful AI solutions in priority sectors including healthcare, agriculture, education (with an emphasis on quality instruction in local languages), public service delivery, and financial inclusion. This fund could consist of an expansion of the recently established 'CurrentAI' foundation.

3. **Develop an international AI data commons with dedicated investment targets**. Establish a publicly accessible repository with high-quality training data for at least 100 languages by 2027. This data commons should be developed with clear governance rules for ethical collection and use while ensuring intellectual property protections.

4. **Launch a comprehensive global AI capacity-building initiative with specific training targets**. For example, this effort could aim to train 100,000 AI practitioners in developing regions by 2028, with specialized tracks for model fine-tuning, open-source development, and safety engineering. This effort could, for example, feature regional training centers offering standardized curricula and certification programs.

4.4.2 Responsible Compute Access

Background

Computing resources represent a critical chokepoint in advanced AI development. Compute governance is an effective lever for influencing the global AI landscape. Access to specialized AI hardware, particularly advanced semiconductors and large-scale compute clusters, is currently concentrated among a small number of nations and companies, creating both security concerns and equity challenges. The governance of compute resources has evolved from general export controls to more AI-specific restrictions, reflecting a growing recognition of compute as a strategic resource with significant security implications.

Balancing security and equitable access creates significant policy challenges. The inherent tension between these priorities is complex: overly restrictive controls could harm economic development in many countries, while insufficient oversight could enable potentially dangerous capabilities to proliferate. Striking the appropriate balance requires international coordination mechanisms that can differentiate between legitimate development needs and security risks, bringing together semiconductor manufacturers, cloud computing providers, nations with advanced chip fabrication capabilities, and countries seeking to develop domestic AI industries.

Recommendations:

1. **Encourage advanced AI states and frontier AI companies to provide frontier-level compute to countries without advanced AI capabilities.** Because developing frontier AI models requires significant energy and data infrastructure—in addition to access to cutting-edge chips in very limited supply—it is unlikely that many countries will develop sovereign, advanced compute abilities. Countries like the US and China should provide compute to these other nations, primarily in the form of computing credits in the short-run and through guaranteed chip deliveries in the long-run. These compute guarantees could come in return for binding guarantees about chip security and guardrails on model development. However, it must be acknowledged that algorithmic and hardware progress may decrease the salience of compute caps in the future, so these security guarantees should be viewed as short-term solutions.

2. **Set in motion longer-term AI development projects, including regional and multi-state AI consortia with dedicated infrastructure funding**. By developing pan-regional development projects, nations that currently lack the ability to develop advanced AI models individually may be able to host robust, joint AI compute capability in the future. However, the physical infrastructure and investment (e.g. power plants, chip factories, data centers) required to make these partnerships possible will take several years to set in motion. Short-term compute agreements as laid out above could address frontier AI needs until longer-term infrastructure projects come online.

3. **Provide compute resources only to countries and developers with robust safety and security mechanisms.** Providing compute resources to third-party countries should come with safety and security stipulations, such as physical and cyber security protocols around data centers, as well as safety and responsible-development requirements for accessing cloud-computing credits. Countries at the forefront of compute development should agree to only supply chips to

third-party countries with these safety and security mechanisms in place, as verified by an independent oversight regime. This regime could consist of an international inspection body tasked with conducting periodic on-site inspections and remote audits to validate infrastructure integrity and adherence to safety and security practices. This regime could also consist of a multinational end-user verification program modeled after financial oversight mechanisms. This program would establish a "white list" of permitted users who have demonstrated that their use and development of AI is strictly for safe purposes. Members of this list would be subject to regular verification procedures to maintain their trusted status.

5. Conclusion

Today's international AI safety landscape leaves significant gaps in addressing and governing severe risks from advanced AI systems. In this report, we offer an overview of the existing landscape, highlight promising opportunities for progress, and propose concrete steps to better integrate, strengthen, and expand current coordination and governance efforts. It is crucial to act now, as AI risks could quickly outpace our ability to respond without concerted global coordination and action. Success depends on early leadership: a small number of states and institutions can already set powerful precedents by formalizing multilateral processes, coordinating scientific assessments, and building safety networks that future governance frameworks will build upon.

As a response, we propose a framework built around four actionable pillars integrating scientific assessment, regulatory standards, institutional coordination, and responsible access. The proposed governance structure would embed scientific expertise into decision-making so that risk assessments directly guide regulatory actions.

While the Bletchley, Seoul, and Paris Summits reflected growing political interest in AI safety governance, these efforts lack institutional continuity, regulatory mechanisms, and structured follow-ups. AI governance must now translate scientific risk assessments into robust policies while keeping regulatory frameworks adaptable to accelerating technological advancements. An independent International AI Safety Research Network (IAISRN) would help keep scientific assessments free from undue political and commercial pressures. AI safety reporting standards would foster transparency into companies' implementation of their commitments. States would back a tiered AI risk classification system that differentiates between Low-Risk, Moderate-Risk, High-Risk, and Unacceptable-Risk AI models. National and regional AI Safety Institutes (AISIs) would conduct localized risk evaluations while maintaining alignment with global safety standards.

Importantly, components of this agenda vary in terms of the level of effort or international consensus required to effect meaningful change. Several proposals could be advanced quickly via leadership from a small group of motivated actors. For example, formalizing the AI Summit Series and strengthening the International Network of AI Safety Institutes (INASI) are feasible near-term steps. Creating the IAISRN and establishing research funds dedicated to AI safety are also achievable by building on existing political momentum. Other areas, like negotiating binding red lines for unacceptable risks or establishing emergency response protocols, will require more sustained international coordination.Nonetheless, this coordination is well worth the effort.

Below, we provide a simplified overview of feasibility and dependencies across key recommendations:

Recommendation	Related Sections/Dependencies	Feasibility
Formalizing the AI Summit Series (4.1.1)	Supports and overlaps with 4.2.1 (Scientific Assessment) and 4.1.2 (INASI coordination). Secretariat and working groups would interact with both. Funding from 4.2.2 could directly fund the secretariat and WGs.	Feasible with moderate effort; requires leadership by one or two anchor countries (for example, as soon as the next Summit in India, December 2025).
Strengthening AISIs and INASI (4.1.2)	Depends on 4.2.1 (Scientific assessment panels) for technical input, and on 4.4.2 (Compute access and safety verification mechanisms).	Relatively feasible because momentum already exists, but needs funding, technical cooperation, and trusted information-sharing.
The Role of the United Nations (4.1.3)	Anchors many things: Scientific Panel (4.2.1), Emergency Protocols and International Treaty (4.3.2), benefit sharing (4.4.1).	Politically harder. Could build progressively (starting with the High-Level Office).
International Scientific Panel (4.2.1)	Will contribute to emergency response mechanisms (4.3.2) and standards for AISIs (4.1.2).	Fairly feasible if hosted within UN structures or tied to INASI. Needs to protect scientific independence.
AI Safety Research Network and Funds (4.2.2)	Could contribute to all recommendations, especially for capacity-building, safe compute access, and independent oversight structures. It could also fund the AI Summit Series secretariat (4.1.1), AISIs (4.1.2), Observatory (4.3.1), and regional safety hubs (linked to 4.4.2 compute access).	Politically hard given funding and coordination needs. Easier to start with philanthropy and the existing authors of the International Report on AI Safety.
Monitoring Corporate Commitments (4.3.1)	Partly depends on expanding independent evaluators under INASI (4.1.2) and ties to compute/resource control (4.4.2).	Possible through voluntary measures and reputational incentives. Harder to make legally binding.

Red Lines and Emergency Response Protocols (4.3.2)	Based on Scientific Panel assessments (4.2.1), Summit Series agreements (4.1.1), and UN emergency mechanisms (4.1.3).	Politically very hard. States fear agreeing to red lines that limit their future AI capabilities. Emergency protocols easier, but still tough.
Responsible Compute Access (4.4.2)	Relies on security inspection mechanisms described in 4.1.2 (AISIs/INASI) and emergency containment procedures in 4.3.2.	Technically feasible; politically harder (esp. if it limits companies' freedom or geopolitical competition intensifies).
Sharing AI Benefits (4.4.1)	Needs funding mechanisms from 4.2.2 (Safety Funds), and political legitimacy supported by the UN office (4.1.3).	Politically sensitive because of redistributive elements. Needs Global South leadership and sustained pressure.

Annex

Regional Approaches to AI Safety

Europe

The European Union broke new ground in AI governance with the EU AI Act, the world's first comprehensive legal framework for artificial intelligence. The Act introduces a risk-based approach that classifies AI systems into three categories (unacceptable, high, and limited risk) and imposes corresponding legal obligations to ensure safety, transparency, and the protection of fundamental rights. Several European countries are also individually taking proactive steps to regulate AI. France is investing in AI ethics and innovation through its National AI Strategy, while Spain has established the Spanish AI Supervision Agency, one of the first national-level AI oversight bodies in Europe. While the Act entered into force in August 2024, it is being applied in phases. From February 2, 2025, the ban on prohibited AI practices (such as social scoring and certain biometric identification systems) and requirements for staff AI literacy took effect. From August 2, 2025, obligations for General-Purpose AI (GPAI) model providers, including documentation, copyright compliance, and data transparency, will apply. GPAI models designated as carrying *systemic risk* will also face additional requirements, including regarding model evaluations, cybersecurity measures, incident reporting, and systemic risk assessment and mitigation. In parallel with the EU's AI Act, the European AI Office is facilitating the drafting of a General-Purpose AI Code of Practice. Chaired by independent experts and involving nearly 1,000 stakeholders, this Code focuses on providers of general-purpose AI (GPAI) models, particularly those that may carry systemic risks due to their advanced capabilities or widespread use. By April 2025, the Code should be finalized, providing a central tool for GPAI model providers to comply with the AI Act's requirements, and the final Code is expected from May 2025 onward. The EU AI Office is also actively developing implementation tools and methodologies alongside the Code.

The Council of Europe (CoE) adopted the Framework Convention on Artificial Intelligence and Human Rights, Democracy, and the Rule of Law on September 5, 2024. Notably, this is the first international binding instrument on AI, and it is open for signature by both CoE Member States and non-Member States. Among its features are a commitment to ensure AI systems respect human rights, uphold democracy, and adhere to the rule of law; a technology-neutral approach intended to remain adaptable to future advancements; and requirements for signatories to implement transparency, accountability, and risk management measures in AI development and deployment. Early signatories include the European Union, the United Kingdom, the United States, and Israel, with the Convention set to enter into force once at least five signatories, including three CoE Member States, have ratified it.

China

China had previously strengthened its role in regional AI governance with the implementation of the Interim Measures for the Management of Generative Artificial Intelligence Services (AI Measures) on August 15, 2023. Under the AI Measures:

- Generative AI models are subject to risk-based oversight, with higher scrutiny for systems capable of influencing public opinion or causing social disruption.
- Providers must ensure lawful data use, protect intellectual property rights, and respect user privacy. Additional measures include transparent labeling, user guidance, and content moderation to prevent harmful or illegal outputs.
- Providers are required to use lawful, diverse, and high-quality data sources, respecting intellectual property and personal information.
- AI systems must uphold "socialist core values", avoid discrimination, and protect user rights, including privacy, reputation, and personal safety.

In 2024, China officially elevated AI safety to the level of national security and public safety, placing it alongside cybersecurity, biological security, and natural disasters. This was emphasized during the Communist Party of China's (CPC) 20th Central Committee's Third Plenum in July 2024. AI providers in China must actively moderate illegal or harmful content generated by their systems and report violations to the Cyberspace Administration of China (CAC), the primary regulatory body overseeing China's AI industry. Regulators also conduct security assessments, inspect service providers, and impose penalties for non-compliance under laws like the Personal Information Protection Law (PIPL) and Data Security Law.

In March 2025, China released the final Measures for Labeling Artificial Intelligence-Generated Content, which will take effect on September 1, 2025. These measures mandate explicit labels (through visible text, audio, or graphics) for AI-generated content that could mislead the public, alongside metadata ("implicit labels") identifying the provider and the content's nature. Online platforms are required to detect these labels, categorize AI content as confirmed, possible, or suspected, and ensure traceability through technical means. In February 2025, the CAC also announced that strengthening AI labeling regulation would be a key task under the 2025 "Qinglang" campaign against online misinformation.

Additionally, China is preparing to implement the Regulation on Network Data Security Management (NDSM) in 2025 and issued draft guidelines in late 2024 for responding to security incidents involving generative AI models. Shanghai and Beijing launched municipal AI safety labs in mid-2024. Over 40 AI safety evaluations have reportedly been conducted by government-backed research centers. China has also been active on the international stage.

United States

AI governance in the United States has changed significantly since the 2024 election. Upon his return to office, President Donald Trump overturned the previous administration's Executive Order on Safe, Secure, and Trustworthy AI issued in October 2023, which had introduced:

- ➢ Requirements for developers of advanced AI systems to share safety test results with the federal government.
- ➢ Measures to address algorithmic discrimination and promote responsible AI use in healthcare and education.
- ➢ Reports on AI's effects on the labor market and strategies to mitigate harm.
- ➢ Expanding bilateral and multilateral AI engagements and supporting international standards for safe AI use.

In January 2025, Executive Order 14179 explicitly revoked the previous AI safety executive order and directed federal agencies to review policies to remove barriers to innovation and ensure AI systems are free from "ideological bias or engineered social agendas." A separate Executive Order on AI Infrastructure, issued January 14, 2025, prioritized national security, economic competitiveness, domestic data center development linked with clean energy initiatives, and workforce development standards for AI-related sectors.

In February 2025, the Office of Management and Budget (OMB) released Memorandum M-25-21, which directed federal agencies to accelerate AI adoption, minimize bureaucratic hurdles, empower agency-level AI leadership, and implement minimum risk management practices for high-impact AI systems.

The 118th US Congress introduced over 40 AI-relevant bills, none of which were enacted as of February 2025. At the State level, California's legislature passed SB 1047 in August 2024, only for the bill to be vetoed one month later after a hefty political debate. SB 1047 attempted to leverage California's role as a hub for major AI companies to address risks associated with frontier models. The bill required AI developers to ensure that their models do not and will not acquire hazardous capabilities such as enabling the creation of weapons of mass destruction, executing large-scale cyberattacks, or causing severe public harm. The bill was vetoed by California's governor, citing concerns about regulatory overreach and potential impacts on innovation, after a pitched debate exposing rifts among AI researchers, technology companies, and policymakers. Without this bill, and after the repeal of the Executive Order, AI companies in America face no AI-specific legally binding oversight, relying instead on voluntary commitments to manage frontier AI risks.

Since then, a new California bill, SB 53, was introduced focusing specifically on whistleblower protections for employees reporting critical AI risks, though this bill is much narrower in scope than SB 1047.

Additionally, the California Frontier AI Working Group issued a draft report recommending measures such as transparency obligations, third-party risk assessments, and adverse event reporting requirements for foundation models. These recommendations are expected to inform upcoming legislative efforts.

The US AI Safety Institute (US AISI) remains active despite the shift in federal policy, continuing to develop testing methodologies, conducting joint model evaluations (such as testing OpenAI's o1 model with the UK AISI in December 2024), releasing draft misuse guidelines for dual-use models (January 2025), and establishing specialized taskforces.


Centre pour
la Sécurité de l'IA