Centre for International
Governance Innovation

CIGI Discussion Paper

# Framework Convention on Global AI Challenges

## Accelerating international cooperation to ensure beneficial, safe and inclusive AI

Duncan Cass-Beggs
Stephen Clare
Dawn Dimowo
Zaheed Kara

## About the Global AI Risks Initiative

The Global AI Risks Initiative at the Centre for International Governance Innovation was created to advance the international governance that will be needed to manage global AI risks. The Initiative aims to mobilize the resources, talent and influence of policy makers, AI researchers, governance experts and civil society to reduce global risks from advanced AI systems. It seeks to build understanding of the importance of global risks from AI and access to workable policy options to mitigate these risks successfully.

This discussion paper proposes the development of an international Framework Convention on Global AI Challenges accompanied by specific supporting protocols to address specific global challenges raised by advanced AI. The paper draws on a wealth of existing research and policy efforts, as well as valuable discussions and feedback from many quarters. Nevertheless, the recommendations are preliminary and intended to support further dialogue, reflection and action. The authors welcome feedback, as well as suggested improvements and collaborations. They may be reached at globalairisks@cigionline.org.

## Credits

Executive Director, Global AI Risks Initiative Duncan Cass-Beggs (lead author)
Senior Research Associate, Global AI Risks Initiative Stephen Clare (lead author)
Program Manager, Global AI Risks Initiative Dawn Dimowo
Senior Research Associate, Global AI Risks Initiative Zaheed Kara
Senior Publications Editor Jennifer Goyder
Publications Editor Susan Bubak
Graphic Designer Sami Chouhdary

The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the Centre for International Governance Innovation or its Board of Directors.

For publications enquiries, please contact publications@cigionline.org.

67 Erb Street West
Waterloo, ON, Canada N2L 6C2
www.cigionline.org

# Table of Contents

# Executive Summary

**Advanced artificial intelligence (AI) could be the most powerful technology ever created by humans,** unleashing explosive growth in cognitive capability that transforms all aspects of society and shapes the future trajectory of civilization. Such a transformation presents unprecedented global challenges. Humanity must realize and distribute global benefits from AI, address global AI risks and make globally legitimate and effective decisions about how to govern advanced AI. This will require the very best of human ingenuity, wisdom and global cooperation over the coming years.

**Future developments in AI could bring enormous global benefits,** with the potential to accelerate scientific discovery, spur technological innovation and massively increase prosperity. However, these benefits cannot be realized and fairly distributed through national policies or market forces alone. International cooperation will be required to enable widespread access to safe AI, harness AI to achieve global public goods and ensure an equitable distribution of the income generated by AI.

**Some global-scale risks from AI can only be managed effectively through international cooperation.** As AI systems become more powerful, they could pose severe safety and security risks worldwide. Such risks may include potential catastrophes such as the intentional misuse of powerful AI systems to cause widespread harm and the loss of human control over autonomous AI systems. Since such risks can cross borders, governments may not be able to ensure the safety of their own citizens unless they cooperate with others.

**International cooperation is also required to enable legitimate and effective decision making on AI developments affecting the future of all humanity.** Currently, a small number of people in a handful of AI companies are making choices that have the potential to affect the lives of people around the world. These choices relate not only to the benefits and risks of AI, but to fundamental questions about whether, and under what conditions, to develop AI systems that vastly surpass human capabilities.

**The international community is not prepared for global AI challenges of this scale.** Important efforts are under way to strengthen international understanding and cooperation on AI, such as through the United Nations and AI Safety Summits. However, these efforts do not yet appear on track to handle some of the most challenging potential scenarios facing the global community, such as the need to detect or prevent the development of unacceptably dangerous AI systems.

**This discussion paper proposes a robust and agile approach to addressing the issues posed by the accelerating development of AI.** This approach consists of swiftly developing and adopting an international Framework Convention on Global AI Challenges, accompanied by specific protocols to facilitate collaborative action on the most urgent issues.

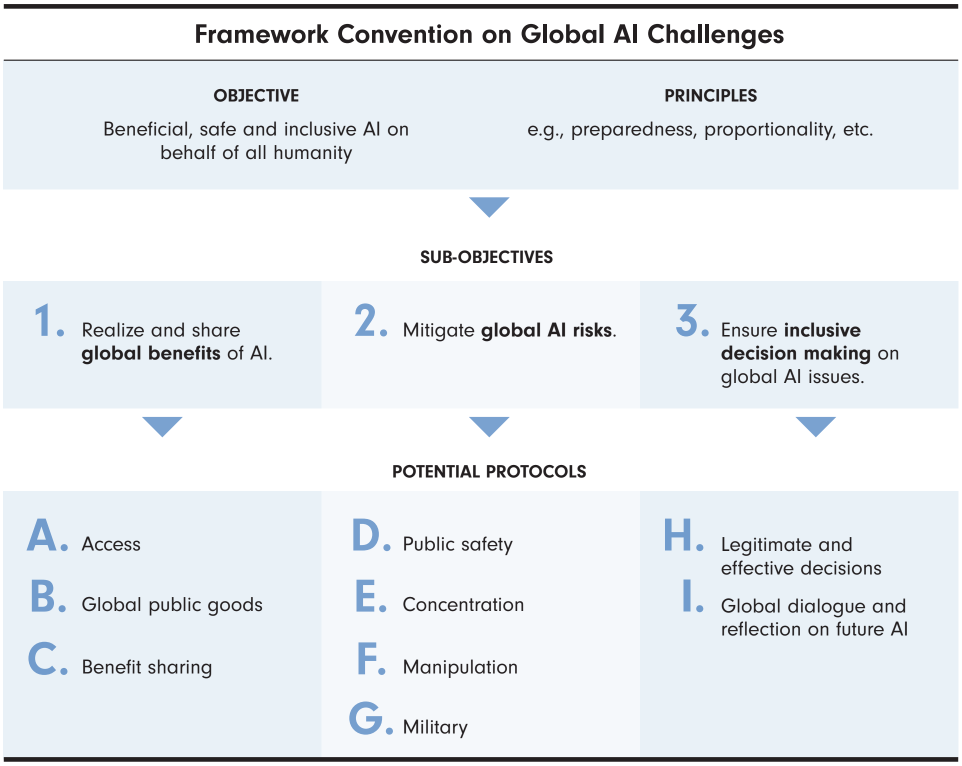
**An international Framework Convention on Global AI Challenges should codify the most important shared objectives and principles** for international AI cooperation, namely, the distribution of global benefits from AI, the mitigation of global risks and the process for legitimate decision making on key choices affecting the future of humanity. The Framework Convention would also set out processes to prioritize and facilitate further international collaboration on the most urgent and important issues of AI governance facing the global community, through the development of timely, specific and effective protocols for joint action.

**A Protocol on Global Public Safety and Security Risks from AI should set out the necessary actions by signatories to successfully mitigate the most urgent and severe global-scale AI risks.** This could include a tiered approach: low-risk AI requires no internationally coordinated regulation; higher-risk AI is subject to coordinated regulation and licensing based on common safety standards; very high-risk AI is only permitted to be developed and run within an international joint AI lab; and AI that poses an extreme risk to humanity is prohibited from being developed until sufficient safety measures can be implemented.

**This Protocol should also establish the necessary instruments for international cooperation** (building on existing bodies where possible) to implement this system effectively. This could include a council to make key political decisions, a commission to provide necessary scientific advice, an agency to set standards and monitor implementation, a lab to conduct joint AI research and an adjudication mechanism to settle disputes. The Protocol could be adopted initially by a core group of parties but should eventually apply universally to ensure that no country develops AI systems that impose unacceptable risks on the rest of humanity.

**The proposed Framework Convention and initial Protocol(s) should be adopted as soon as possible** and improved incrementally over time, given the rapid pace of AI development and the (albeit small) possibility that AI systems posing global-scale risks could be developed within the next few years.

**Next steps in advancing this work include** engaging with diverse global stakeholders, including states, to develop and refine a model Framework Convention and Protocol(s), stress-testing proposed measures under different scenarios, integrating recommendations into official international negotiation processes and expanding research to fill critical knowledge gaps.



**Framework Convention on Global AI Challenges**

OBJECTIVE

Beneficial, safe and inclusive AI on behalf of all humanity

PRINCIPLES

e.g., preparedness, proportionality, etc.

SUB-OBJECTIVES

1. Realize and share **global benefits** of AI.
2. Mitigate **global AI risks**.
3. Ensure **inclusive decision making** on global AI issues.

POTENTIAL PROTOCOLS

A. Access
B. Global public goods
C. Benefit sharing

D. Public safety
E. Concentration
F. Manipulation
G. Military

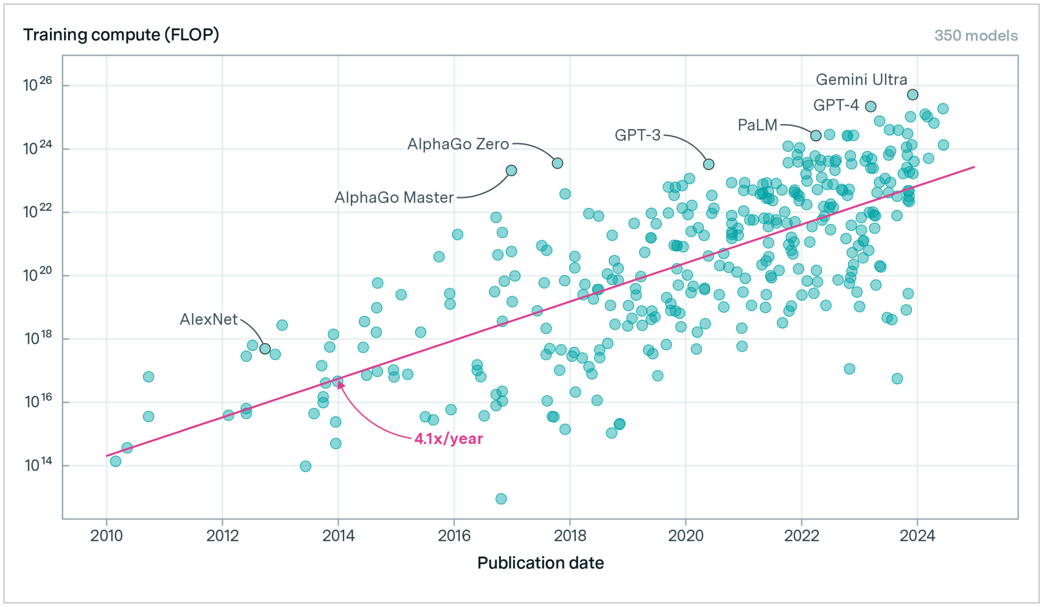
H. Legitimate and effective decisions
I. Global dialogue and reflection on future AI

# Introduction

**Humanity may still be vastly underestimating AI.** Recent advances in AI are garnering widespread attention, with an accelerating pace of breakthroughs and releases matched by a flourish of media commentary, public debate and national and international policy initiatives. However, most people are still largely focused on the types of AI systems that exist today, while significantly underestimating and underpreparing for the potential power, pace of development and scale of implications of AI systems that could exist in the years ahead.

**AI is currently on track to far surpass humans in terms of the power to interpret and act upon the world.** In recent years, AI systems have reached or surpassed human-level performance across an increasingly wide range of cognitive tasks, such as reading comprehension and language and image generation (Giattino et al. 2023). These developments have been driven by large increases in the amount of computing power (see Figure 1) and in the size of data sets, and improvements in the quality of algorithms used to train AI systems. If these trends continue, the capabilities of AI systems could eventually reach and vastly exceed human performance in most or all cognitive domains.[1] Such cognitive proficiency could then enable interaction with other systems, as well as increased capabilities for physical action in the world, such as through rapid advances in robotics.

Figure 1: Computing Power Used to Train Notable AI Models



*Note:* This graph displays the trend in computing power (compute) used to train large AI models. FLOP = floating-point operations.

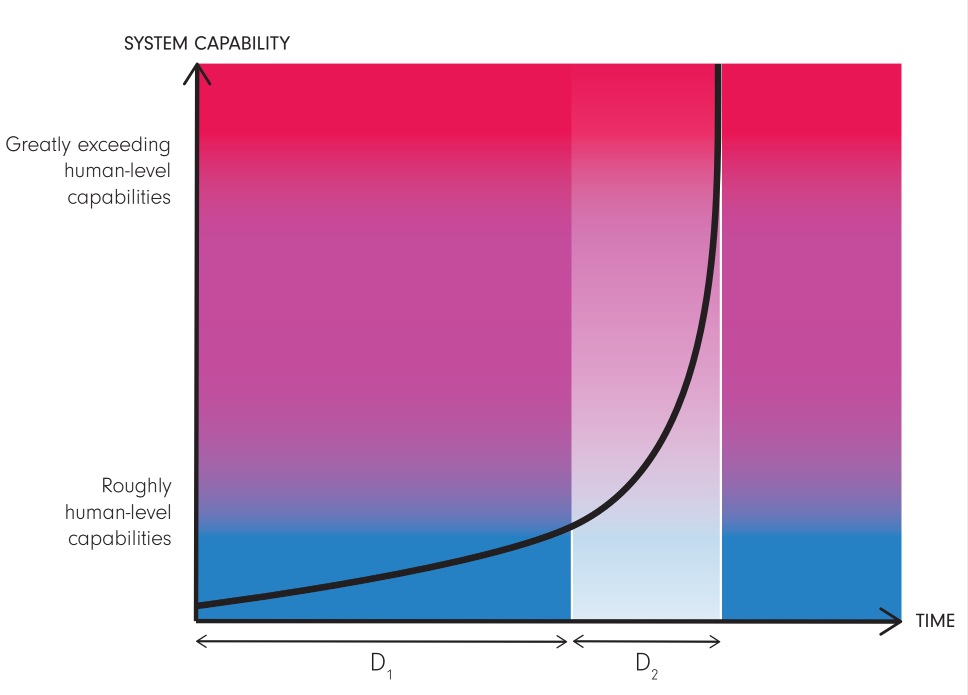
*Source:* Sevilla and Roldán (2024).

**Although the pace of AI development is highly uncertain, there are reasons to believe that these advances could occur rapidly** (Bengio et al. 2024). The world's largest companies are investing billions of dollars in AI development, leveraging top talent and computing resources to overcome technical obstacles and unlock additional jumps in capabilities. Automated AI research assistants may also tangibly assist in AI development, resulting in a feedback loop of AI improvements enabling further

[1] An AI system capable of human-level performance on most or all cognitive tasks is sometimes referred to as *artificial general intelligence* (AGI). A system that would be capable of greatly outperforming humans on most or all cognitive tasks is sometimes called *artificial superintelligence* (ASI). For a more detailed exploration of definitions, see, for example, Morris et al. (2024) and Maas (2023).

AI improvements. Furthermore, AI models typically require many times more computing power to train than to run. Therefore, after one capable AI assistant has been trained, many instances of it can be run, working constantly and in concert to accelerate AI development yet further. In sum, it is possible that AI progress is nearing the very steep part of an exponential curve, enabling increasingly powerful AI systems (Everitt, Lea and Hutter 2018; see Figure 2).

Figure 2: A Hypothetical Graph of AI Capabilities over Time



*Note:* A graph of AI system capabilities over time, depicting a hypothetical scenario in which AI systems assist with AI research. This initiates a positive feedback loop enabling the rate of capabilities development to become exponential. In this case, the length of time ($D_1$) to reach approximately human-level capabilities would be longer than the length of time ($D_2$) to greatly exceed human-level capabilities.

*Source:* CIGI. Adapted from a similar chart shared by the International Center for Future Generations.

**AI developments in the next few years could have implications far beyond what nearly anyone is currently anticipating or prepared for.** Increasingly powerful AI could enable prosperity great enough to erase economic inequalities, or massively amplify them. It could transform the power balance within and across societies. It could bring about game-changing solutions to some of the biggest global challenges, such as climate change, while generating new dangers on a global scale. It could replace humans in an increasing range of core functions, ultimately threatening to supersede humans as the dominant societal actors. Humanity may therefore face choices of unprecedented magnitude and consequence about what kinds of AI to create, for what purposes and under what conditions.

**This discussion paper aims to inform discussion on how the international community can best prepare for and manage the potential implications of advanced AI.** The paper begins by exploring three emerging global-scale challenges posed by advanced AI that could require international cooperation. These relate to realizing the global benefits of AI, mitigating the global risks and making legitimate choices about future implications of AI for humanity. The paper then recommends an instrument to help facilitate international cooperation on these issues: a Framework Convention on Global AI Challenges. Such a framework convention would have the breadth and flexibility needed to respond to the urgent and transformative nature of global AI challenges. Finally, the paper recommends prioritizing the adoption of a specific Protocol on Global Public Safety and Security Risks from AI, given the potential severity and uncertain timelines of such risks.

# Global Challenges from Advanced AI

**International cooperation will be needed to manage emerging global challenges arising from the development of advanced AI.** While many AI issues can be managed at the local or national levels, some have substantial cross-border effects and cannot be addressed effectively or legitimately by one government alone. To govern these issues, states will likely need to come together to negotiate and implement some form of international agreement.

**This paper examines three global AI challenges in particular which humanity could face in the coming years, and which likely require international cooperation to overcome.** These are:

- how to realize and share the **global benefits** of AI;
- how to mitigate severe **global risks** posed by AI; and
- how to make **legitimate and effective decisions** about the future implications of AI for humanity.

## Challenge One: Realizing and Sharing Global Benefits of AI

**Advanced AI systems could bring about considerable benefits.** They are expected to accelerate economic growth by complementing and substituting for human labour and accelerating scientific advances. This increased productivity could be harnessed to help solve some of the largest challenges humanity faces today. If applied effectively and equitably, AI systems could help alleviate poverty, cure diseases, invent green technologies and empower people around the world with tailored knowledge and advice. AI-driven productivity growth could also significantly increase the quantity and quality of material goods such as food, transportation, energy and housing, as well as intangible goods in the form of education, entertainment, support and companionship.

**However, many of these positive outcomes will likely not be achieved by market forces alone.** Government action, and international cooperation in particular, may be required to address three likely shortcomings: the underutilization of AI systems due to lack of access; the underprovision of global public goods from AI; and unequal distribution of AI benefits.

**First, global cooperation may be necessary to ensure that all people are able to access productivity-enhancing AI technologies.** People in lower-income countries are already grappling with digital infrastructure deficits, prohibitive costs of connectivity and barriers to developing digital skills. As a result, they are unlikely to access and adopt AI tools and services at the same rate as their counterparts in advanced economies. For example, according to figures from the International Telecommunication Union (2023), just 37 percent of people in Africa use the internet, compared to up to 91 percent in Europe. Even within wealthy countries, there is a risk that access to the most advanced technologies may be concentrated among a small number of leading firms and highly skilled individuals. The lack of access to AI tools has global-scale costs in terms of both inequality and lost productivity. International cooperation could address this shortcoming by prioritizing wide-scale access to safe, secure and trustworthy AI.

**Second, international cooperation may be needed to harness the power of AI to produce global public goods.** These are products and services that provide significant global benefits, but which private companies and individual governments lack sufficient incentives to provide. Such goods could include educational resources, vaccines, green technologies and other useful technologies. The international community could coordinate to produce these goods by subsidizing the use of advanced AI systems to provide them.

**Third, international cooperation will be required to make the distribution of AI benefits more equitable.** Some of the benefits of an AI-driven productivity surge may be widely distributed through lower prices. However, a large share of the surplus generated by AI could accrue to a limited number of individuals or organizations controlling the most productive AI systems and associated value chains. International cooperation could address this issue by designing tools to fairly tax and redistribute benefits. Designing, implementing and enforcing mechanisms to do so is a considerable challenge, but could be accomplished through, for example, various types of income transfers or subsidized production of key goods and services.

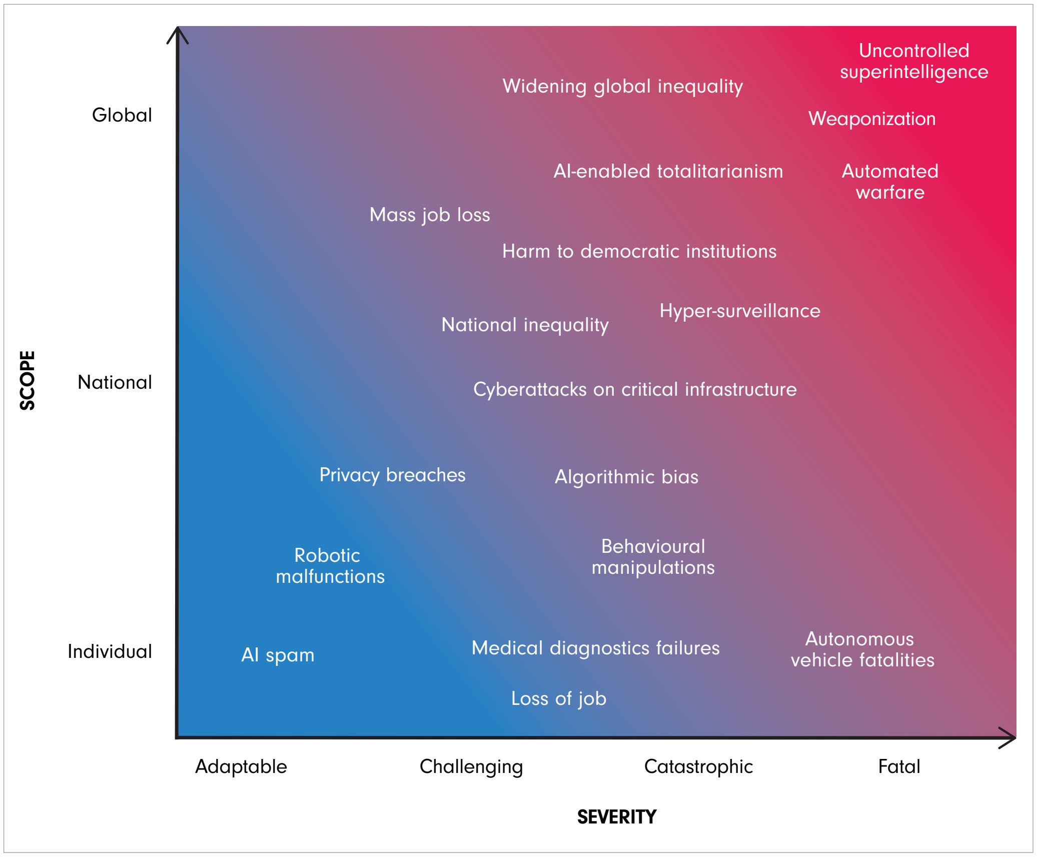
**The productivity gains enabled by advanced AI are potentially large enough to support far more global income redistribution than has been politically feasible in the past.** Commitments to such redistribution are more likely to be accepted and honoured if they are made in advance, before the gains have been realized. This principle is reflected, for example, in the proposed Windfall Clause, whereby AI companies would commit to donate all profits above a certain level of income (for example, one percent of global economic output) (O'Keefe et al. 2020). Ensuring that all of humanity benefits from AI is likely also a moral and political imperative given that the most severe risks from AI affect all of humanity equally.

## Challenge Two: Addressing Global AI Risks

As a rapidly developing technology with multiple potential capabilities and uses, advanced AI systems pose numerous risks. A number of these risks are illustrated in the chart below (see Figure 3)[2] and described in more detail in Appendix 1. Some of the risks are familiar, such as issues of privacy, bias, misinformation and labour market risks (GOV.UK 2024a). But experts have also raised concerns that AI could pose catastrophic risks (Anderljung et al. 2023; Bengio et al. 2024; Critch and Russell 2023). Although such extreme outcomes may seem unlikely, they cannot currently be ruled out.

[2] The positioning of risks on the chart is intended to be illustrative and does not claim to be comprehensive or authoritative. The authors welcome feedback, including rationales for alternative positioning as well as suggestions of additional risks.

Figure 3: Scope and Severity of AI Risks



*Note:* AI risks are plotted in two dimensions, scope and severity. Scope refers to the range of a risk's effects and ranges from affecting individuals to affecting the whole world. Severity refers to the seriousness of a risk's effects and ranges from adaptable, which may be imperceptible or easily mitigated, to fatal, causing death or irrecoverable damage.

*Source:* CIGI.

**In the worst case, AI risks could be both global in scope and extreme in severity.** This implies they could cause enormous economic damage or loss of life at the global level. How might such devastating outcomes occur? Here the focus will be on two types of risks of this nature: weaponization and loss of control.[3]

[3] These are not the only possible global catastrophic risks from AI. For example, mass adoption of AI systems and widespread automation of decision making creates new systemic vulnerabilities that could result in sudden collapse. Loss of control could also result not from sudden AI takeover, but rather from humans being gradually and systematically outcompeted by AIs that can perform their jobs faster, cheaper and more reliably. Another suggested variety of existential threat is one where humanity may not be destroyed or replaced but could suffer values lock-in due to the power of human-controlled or autonomous AI to shape human beliefs and prevent further developments in human civilization. This paper focuses on weaponization and loss of control as two particularly clear and important risks.

## AI Weaponization

**AI weaponization concerns the misuse of AI systems to cause harm.** Experts have expressed concern, for example, that threat actors could use advanced AI technologies and advanced biotechnologies in tandem to engineer novel pathogens and unleash a deadly pandemic, possibly on a global scale (Sandbrink 2023; Urbina et al. 2022). AI may also aid in the development of potent cyberweapons capable of disrupting critical infrastructure globally (Shevlane et al. 2023). Generative AI systems could allow misinformation to spread at scale to provoke widespread societal upheaval, violence and conflict (Matz et al. 2024; Shevlane et al. 2023). And as AI tools become even more powerful, they could assist malicious actors in developing new hazards and strategies for widespread harm that humanity has not yet prepared for.

**The challenge is exacerbated by AI broadening the range of actors capable of causing severe harm**, potentially empowering terrorist groups, rogue states, organized crime units, and incautious organizations and individuals to develop and deploy such weapons. Although AI will also help defend against such threats, it is not clear that defensive improvements will consistently keep pace with new offensive capabilities (Garfinkel and Dafoe 2021), or that these defences (such as mass surveillance and automated cyber defences) will not pose global-scale risks of their own (Peterson and Hoffman 2022).

## Loss of Control

**Another category of AI risks concerns the loss of control of advanced AI systems.** As AI systems become more capable, there are increasing efforts to develop autonomous AI agents that can make plans and pursue goals with limited oversight (GOV.UK 2024a). AI agents could prove highly useful, but also more difficult to understand and control. Deactivating them could also be problematic if, for example, they are integrated into important systems such as financial markets or can replicate themselves faster than operators are able to shut them down.

**Scenarios of extreme loss of control could involve the *purposeful* elimination or disempowerment of humanity by a hostile or misaligned AI or group of AIs.** Some AI systems, if their goals are not properly specified or their actions not appropriately limited, may seek to deceive their operators (Hubinger et al. 2024), gather resources and evade shutdown attempts. In the extreme, an AI system exhibiting such misalignment could seek to shut humanity out of critical global systems such as digital infrastructure or military command structures. Alternatively, it may seek to eliminate humanity outright. Such outcomes may seem unrealistic, but no techniques currently exist to reliably align an AI's behaviour with complex or even common-sense human values (Bengio et al. 2024). Autonomous AI systems may pursue goals that are destructive to humanity if such outcomes are seen as compatible with or desirable for its learned or programmed objectives. Sufficiently powerful and intelligent systems may cause severe global catastrophes as a result.

**While there remains high uncertainty about the timing and likelihood of such risks, given their potential scale and severity, immediate safeguards are warranted.** Recent trends suggest that even the next generation of AI systems could have dangerous capabilities (Fang et al. 2024; OpenAI 2023). Especially since capabilities can develop in surprising and hard-to-predict ways (Bowman 2023), this uncertainty calls for a high degree of caution. International cooperation could prove critical to design and implement mechanisms to monitor the development of potentially dangerous AI systems globally, identify systems deemed too dangerous to develop and release, prevent the development of such systems if necessary, and ensure future AI research and development proceeds safely.

## Challenge Three: Making Globally Legitimate and Effective Decisions about How to Govern Advanced AI

**The final global challenge posed by advanced AI is that of making legitimate and effective global decisions** about how to manage the transformative benefits, catastrophic risks and other fundamental issues AI raises for the future of humanity. Currently, decisions about developing and releasing new AI systems are being made by a small and unrepresentative group of people in high-income nations who do not have the legitimacy to make these decisions for the world at large. Such decisions plausibly impose severe risks on people around the globe and should be made by processes that appropriately represent the interests of all affected parties.

**The prospect of advanced AI poses many fundamental ethical questions for all of humanity.** If large swathes of the economy are automated and economic and scientific progress accelerates, humanity may find itself asking deep questions about how to ensure global prosperity, prioritize problems to solve, manage major risks and navigate other issues resulting from the societal impacts of advanced AI, such as space exploration and resource exploitation resulting from new technologies. Humanity will also need to decide on such questions as when to delegate oversight of critical systems to AI advisers and agents. What safeguards should be put in place? If different nations make different choices, how will the potential shifts in global power balances be managed? And if advanced AI systems gain moral and legal standing themselves, what rights will they have (Sebo and Long 2023)?

**Given the complexity of these questions, and the possibility of rapid developments in AI, inclusive processes of reflection and deliberation are needed now.** Such efforts, and the structures and institutions that enable them, could be designed so that whenever humanity faces such questions, its answers are determined by an effective, just, representative and legitimate process.

# Current International Cooperation Efforts Are Insufficient for Global AI Challenges

**International cooperation on AI issues has accelerated with the pace of AI development in recent years.** These efforts demonstrate an encouraging resolve among governments and multilateral organizations to cooperate to address AI challenges. The UN resolution on safe, secure and trustworthy AI, adopted unanimously by the UN General Assembly in March 2024, covers a broad range of issues relating to both the benefits and risks of AI (United Nations General Assembly 2024). The UN Secretary-General has also convened a High-level Advisory Body on Artificial Intelligence to develop recommendations for globally coordinated AI governance, and in September 2024 the UN Summit of the Future will aim to finalize a Global Digital Compact that outlines shared global values and principles for AI (Advisory Body on Artificial Intelligence 2023; United Nations 2021). Significant international cooperation efforts on AI are also being pursued by the G7, the Organisation for Economic Co-operation and Development (OECD), the African Union, the Council of Europe, the European Union, the UK AI Safety Summit follow-up process, as well as through various formal and informal bi- and "mini-lateral" discussions (see Box 1 for details).

### Box 1: International Coordination Efforts on AI Issues

In addition to work being conducted under the aegis of the United Nations, a number of other international coordination efforts on AI issues are under way. For example:

- The **G7** plans to advance the Hiroshima AI Process by building out the mechanisms to monitor voluntary compliance with its Code of Conduct for Organizations Developing Advanced AI Systems in collaboration with the OECD and other partners. The code calls on organizations developing AI to take appropriate and commensurate measures to identify, evaluate and mitigate potential safety, security and societal risks (G7 2023).

- The **African Union** is finalizing a Continental Strategy on AI, which has the dual objectives of promoting AI adoption and addressing its associated risks (African Union 2024).

- **China** is championing an inclusive approach, announcing a Global Governance Initiative (Ministry of Foreign Affairs, People's Republic of China 2023) and expressing support for a UN-led process to establish an international AI governance institution.

- The **Council of Europe's** Framework Convention on AI and Human Rights will be open to all countries to join and potentially applies to the development of frontier AI models, though each country can decide on its application to the private sector (Council of Europe 2024).

- The United Kingdom convened an international **AI Safety Summit**, which culminated in the Bletchley Declaration by the European Union and 28 countries affirming the need for the safe development of AI, including the urgent need to better understand risks from frontier models and how to address them (GOV.UK 2023). The attendees of the UK AI Safety Summit also agreed to support a scientific consensus-building report on the "safety of advanced AI," and the interim report was published in May 2024 (GOV.UK 2024a).

- The subsequent **Seoul AI Summit** saw countries commit to collaborate on developing common thresholds for severe AI risks and to launch a network of national AI safety institutes, which could help build consensus around policies and norms (GOV.UK 2024b; 2024c).

**However, these efforts are currently insufficient for managing possible scenarios in which AI systems have rapid, transformative effects around the world**. There are no adequate measures in place to realize and distribute the scale of potential benefits that advanced AI could make possible. There is also a lack of measures to prevent or mitigate newly emerging global-scale risks from AI, such as consistent standards and mechanisms to ensure AI development proceeds safely everywhere (Cihon 2019; Trager et al. 2023), and monitoring and enforcement capacity to ensure universal compliance with them (Heim et al. 2024). Finally, there is a lack of effective and legitimate processes to make decisions regarding AI that could have lasting implications for humanity, such as decisions about when and under what conditions to develop AI systems vastly more capable than humans.

**In summary, given the potential power and pace of development of advanced AI systems, the significant global challenges they may pose, uncertainty about when such challenges may arise, and the apparent inadequacy of current systems to handle these challenges, an effective, future-ready approach to international cooperation on AI governance is urgently required.**

# A Framework Convention on Global AI Challenges

**Successfully navigating emerging global AI challenges is likely to require unprecedented international agreement and cooperation.**[4] Since advanced AI systems, regardless of where they are developed, can have significant implications for other parties, states will likely only be able to secure their interests by cooperating with others. Such cooperation will likely require various verifiable and enforceable international agreements to ensure that all parties meet their commitments. This is likely to be particularly important given the strong incentives for states to, otherwise, seek advantage by breaking or ignoring joint commitments (Horowitz 2018).

**A Framework Convention on Global AI Challenges could provide a practical and flexible instrument to help accelerate international cooperation on the effective governance of advanced AI.** A framework convention[5] can be broad and flexible, allowing the international community to act first on pressing AI challenges while recognizing the full range of issues that must be addressed. When signing a framework convention, parties agree on shared objectives and principles, as well as on processes for addressing contentious issues in accompanying protocols that tackle specific issues and commitments. Those protocols can then be negotiated and signed individually. Early protocols on urgent issues can thus be prioritized while allowing additional time for other protocols that may be less urgent or more contentious.

**The following sections outline preliminary suggestions regarding possible key components of a Framework Convention on Global AI Challenges, and a supporting Protocol on AI Safety.** Given the scope and complexity of these issues, many questions remain to be answered through further research and negotiation.

## Contents of the Framework Convention

### Objective

**The framework convention should establish a clear, high-level objective, such as "ensuring the development of beneficial, safe and inclusive artificial intelligence on behalf of all humanity."** Such a goal is broad enough to encompass a range of sub-objectives aimed at tackling each of the three global challenges of advanced AI, namely: promoting and distributing the global benefits from AI; preventing and mitigating global AI risks; and strengthening global cooperation to ensure legitimate and effective decision making on current and future AI issues.

[4] Global AI challenges are arguably unprecedented in terms of the possible scale and significance of their implications, the limited margin for error, the strong private and national incentives to default on joint commitments, and various practical obstacles to verifying and enforcing compliance. Nevertheless, many other areas of international cooperation can serve as partial models and inspiration, such as nuclear energy and nuclear non-proliferation, climate change, aviation, financial institutions and others.

[5] A framework convention is a kind of international treaty. Examples include the UN Convention on Climate Change (https://unfccc.int/process-and-meetings/what-is-the-united-nations-framework-convention-on-climate-change) and the UN Convention against Transnational Organized Crime (www.unodc.org/unodc/en/organized-crime/intro/UNTOC.html).

## Principles

**The principles of the framework convention should establish a shared foundation of values, priorities and commitments to guide participating states in their negotiations and decision-making processes.** Principles should reflect key shared values such as cooperation, inclusivity, equity, proportionality, effectiveness, preparedness and adaptability.

More specifically, and building on recent work, the principles could:

- recognize the necessity of international coordination and **cooperation** in addressing the global opportunities and challenges presented by advanced AI;
- affirm the need for **inclusive** decision making to ensure that global benefits and global risks accrue and are borne **equitably** across humanity;
- assert that the extent of international cooperation should be **proportionate** to the scale of the global challenges involved;
- assert that joint action must be sufficient to be fully **effective** in achieving the shared objectives set out in the framework convention, while at the same time **limited** to infringe as little as possible on state sovereignty and other goals;
- commit to addressing the wide range of global challenges that AI presents, while **prioritizing** work on the most urgent ones, such as critical global risks to public safety; and
- recognize that, given the fast-changing nature of AI and uncertainty about future opportunities and risks, joint action should be **prepared** to handle the worst-case scenarios, while remaining readily **adaptable** to a wide range of possible future conditions.

## Organizing Bodies

**The framework convention should establish any institutional bodies necessary to convene and guide future discussion on specific protocols and commitments**, beginning with a "Conference of the Parties" of signatory states. Additional institutional bodies may be developed as required to implement, monitor and enforce specific governance measures. In the interest of reaching swift agreement, however, it may be advisable to reserve the establishment of such additional institutions for subsequent protocols.

## Other Components

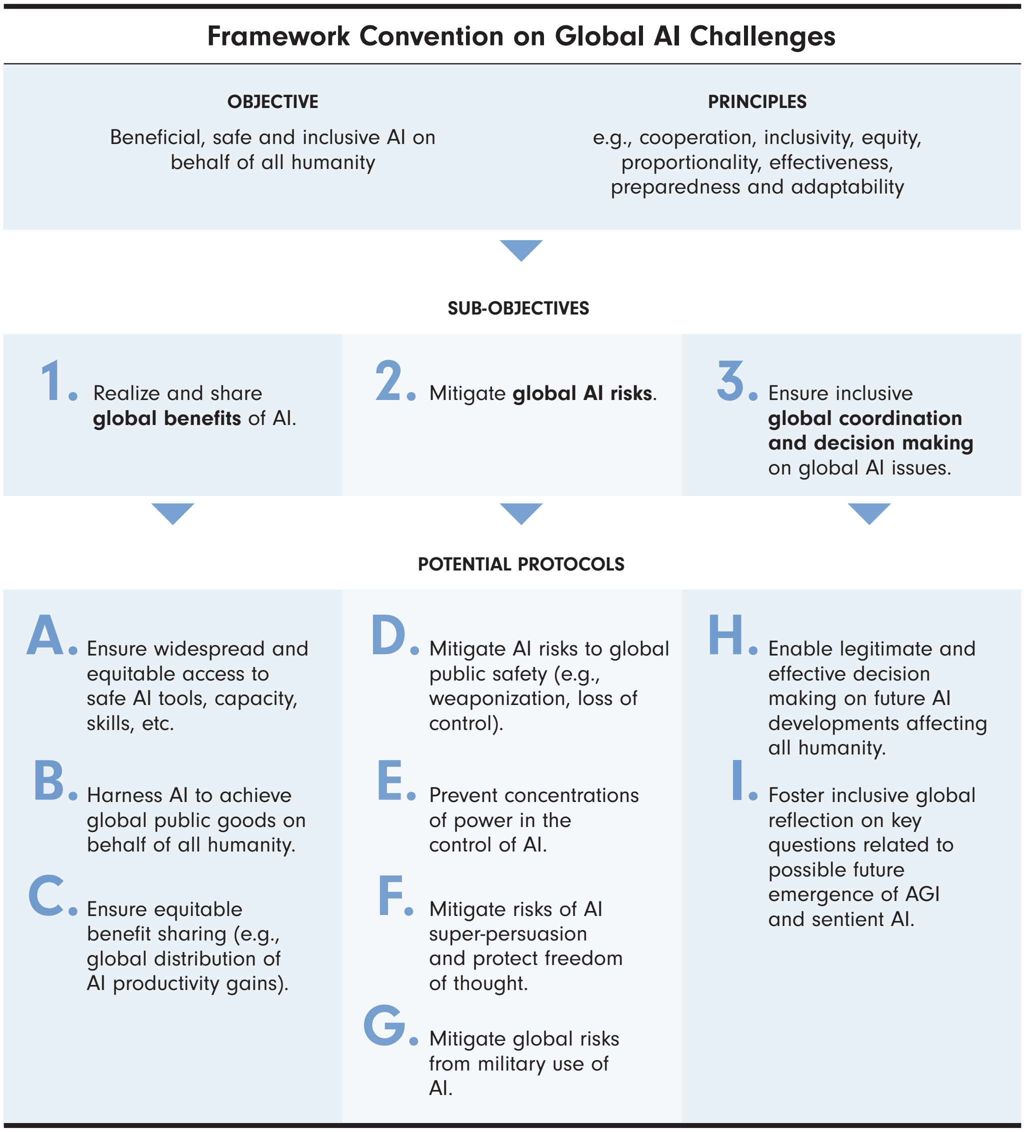
**Framework conventions typically include several additional components to aid future dialogue among participatory states.** These include mechanisms to review implementation, promote compliance and resolve disputes, along with other implementing clauses.

After these structural components are established, the framework convention could be signed, allowing parties to develop supporting individual protocols with more specific commitments.

## Illustration of Framework Convention and Protocols

**Figure 4 provides an overview of the proposed framework convention structure and its relationship to the accompanying protocols.** The examples of potential protocol topics are illustrative only. Some may already be partially covered by existing international agreements or be better suited to separate legal instruments. They have been roughly grouped according to which of the three global AI challenges they help address.

Figure 4: International Framework Convention on Global AI Challenges



**Framework Convention on Global AI Challenges**

**OBJECTIVE**

Beneficial, safe and inclusive AI on behalf of all humanity

**PRINCIPLES**

e.g., cooperation, inclusivity, equity, proportionality, effectiveness, preparedness and adaptability

**SUB-OBJECTIVES**

1. Realize and share **global benefits** of AI.

2. Mitigate **global AI risks**.

3. Ensure inclusive **global coordination and decision making** on global AI issues.

**POTENTIAL PROTOCOLS**

A. Ensure widespread and equitable access to safe AI tools, capacity, skills, etc.

B. Harness AI to achieve global public goods on behalf of all humanity.

C. Ensure equitable benefit sharing (e.g., global distribution of AI productivity gains).

D. Mitigate AI risks to global public safety (e.g., weaponization, loss of control).

E. Prevent concentrations of power in the control of AI.

F. Mitigate risks of AI super-persuasion and protect freedom of thought.

G. Mitigate global risks from military use of AI.

H. Enable legitimate and effective decision making on future AI developments affecting all humanity.

I. Foster inclusive global reflection on key questions related to possible future emergence of AGI and sentient AI.

# A Protocol on Global Public Safety and Security Risks from AI ("Protocol on AI Safety")

**An early priority under the Framework Convention on Global AI Challenges should be to negotiate a protocol focused on reducing risks of catastrophic harm from advanced AI.** This suggested prioritization is based on the following considerations:

- **The potential severity of the harms.** Weaponization or loss of control of AI could potentially cause a catastrophic scale of harm, possibly including trillions of dollars in economic damage and millions or even billions of human deaths.

- **The possibility of short timelines.** While there remains high uncertainty about the potential timing and likelihood of global-scale AI risks to public safety, there is reason to believe that such harms could be possible within even just a few years, or less. This makes acting swiftly to begin reducing these risks a priority.

- **A shared interest in success.** Societies share a common interest in avoiding catastrophic suffering and loss of life. Despite the significant challenges of achieving meaningful international cooperation on AI, agreement on focused, shared areas of interest such as safety and security may be somewhat easier to achieve, whereas agreement on other global AI challenges may take more time.

Prioritizing swift progress on a Protocol on AI Safety does not imply that this protocol should take precedence over other areas. Governments have the capacity to pursue multiple international cooperation goals in parallel, and it is possible that other global challenges relating to AI merit similarly rapid attention and progress. However, global AI catastrophes must be avoided if humanity is to realize the benefits of AI and have the opportunity to make long-lasting decisions about its future trajectory.

## Contents of the Protocol on AI Safety

### Objective and Principles

**The overarching objective of this Protocol is to enable governments and other relevant parties to collaborate effectively to prevent and mitigate severe AI risks to global public safety.** Such risks include the possible misuse or weaponization of powerful AI systems, or the potential loss of human control over AI, such as through the accidental or intentional development of an uncontrolled artificial superintelligence.

The principles of this Protocol could include:

- Any global or globally coordinated restrictions on AI shall be proportionate to the level of risk (that is, likelihood times severity of impact) of the AI system in question.

- Any global or globally coordinated restrictions on the development or use of AI shall be designed to ensure reliable safety from global catastrophic risk (with an adequate margin of error) while otherwise minimizing negative impacts of such restrictions for other global priorities such as innovation, prosperity, privacy, liberty and so on.

- Global risks from AI should only be accepted where they are sufficiently justified by global benefits from AI.

- All decisions relating to global-scale risks (that is, risks borne by all of humanity) should be made by a legitimate and globally representative process.

- Consideration for the interests and rights of future generations.

## Obligations

**Signatories to this protocol, recognizing the importance of reducing global risks to public safety from AI systems to acceptable levels, would mutually commit to several obligations.** More research, as well as dialogue among global stakeholders, is necessary to determine which specific obligations should be adopted. Monitoring and enforcement procedures necessary to enforce those obligations also require additional development. However, some possible obligations that could be considered for the protocol include:

- No state shall allow in its sphere of effective control the development or use of an AI system that poses an unacceptable risk to global public safety (unless authorized to do so by the international community).

- States shall abide by and enforce within their sphere of effective control those standards and policies established by the international community to manage global risks to public safety, such as risk tolerance thresholds, risk management frameworks, regulations applying to compute providers and AI model developers, liability regimes, the accreditation of third-party safety evaluators, and so forth.

- States shall fully cooperate with verification efforts required by the international community to ensure full compliance with the provisions of the Protocol, such as through appropriate monitoring of relevant AI facilities within each state's sphere of effective control.

## Risk Tolerance Thresholds

**A central principle of this Protocol is that governance and regulation of AI systems should be proportionate[6] to the level of global benefits and global risks associated with such systems.** The aim is to ensure an adequate level of safety (for example, a risk of global catastrophic harm that is as low as reasonably practicable) without sacrificing humanity's ability to reap the many benefits of AI or pursue other global priorities such as human rights and UN Sustainable Development Goals.

**A tiered, risk-based approach to the regulation of AI would apply a different level of global restriction based on the assessed risk level of the AI model or system in question (Koessler, Schuett and Anderljung 2024).** An illustrative four-tier system of potential risk tolerance thresholds for existing or proposed AI models and systems, and their corresponding implications in terms of globally coordinated governance restrictions, would comprise:

- → **Tier one: Negligible risk** to global public safety. No common or coordinated global restrictions required. Individual governments or multilateral coalitions may regulate such systems or not as they desire. Globally harmonized liability frameworks may still apply.

- → **Tier two: Manageable risk.** All entities involved in the development or use of such AI systems are required to have a licence. The licensing regime could be nationally administered but must adhere to global common standards. Licensing requirements could include such elements as third-party safety evaluations and demonstrating sufficient standards of cybersecurity to secure model weights from potential leaking or hacking by determined adversaries.

- → **Tier three: Tolerable risk.** AI systems assessed at this level of risk would be permitted to be developed and run exclusively in an international joint AI lab. The purpose of this restriction is to

[6] Note that versions of this principle are reflected in numerous national and international texts on AI governance, including the Canadian draft legislation Bill C-27 and Voluntary Code of Conduct, and in the 2024 EU AI Act.

avoid the race dynamics involved in competition between companies or countries that can result in cutting corners on safety.

→ **Tier four: Unacceptable risk.** AI systems assessed at this level of risk would not be permitted to be developed anywhere until adequate safety and control mechanisms become available to lower the risks to tolerable levels.

**This tiered approach encourages political debate to focus on the level of risk that society is prepared to tolerate.** Such decisions could be made with reference to examples from other safety-critical industries, such as aviation, nuclear energy, transportation infrastructure, pharmaceuticals or other analogous sectors. For example, the chance of severe core damage in nuclear plants is typically required to be less than one in 10,000 per year (International Nuclear Safety Advisory Group 1999). Such discussions could also consider the likelihood and value of potential global benefits from the same AI systems, such as in saving lives through accelerated medical advances, improved prosperity or helping society to address other catastrophic risks such as biological risks or climate change.

**This approach leaves the technical question of the assessed risk of a given system (that is, the correspondence between the system's capabilities and the risk it generates) to be answered by the best available science.** These methods are still very much in their infancy and include evaluations of dangerous capabilities, and the association of potential capabilities with various parameters and features of the AI systems in question, such as the amount of computing power used in training (for example, in floating point operations), model size and so on.

**In the absence of scientific consensus or certainty regarding the actual level of global risk posed by a certain type of AI system, a precautionary approach is likely advisable.** A reasonable range of risk estimates should be generated and regulatory decisions should be informed by the most conservative ones. This would motivate developers to improve the science of risk evaluation. Although assessing such risks seems difficult now, the science of interpretability and evaluation is advancing rapidly. It is possible that reliable, transparent and reproducible methods of evaluating model risk will be developed in the coming years.

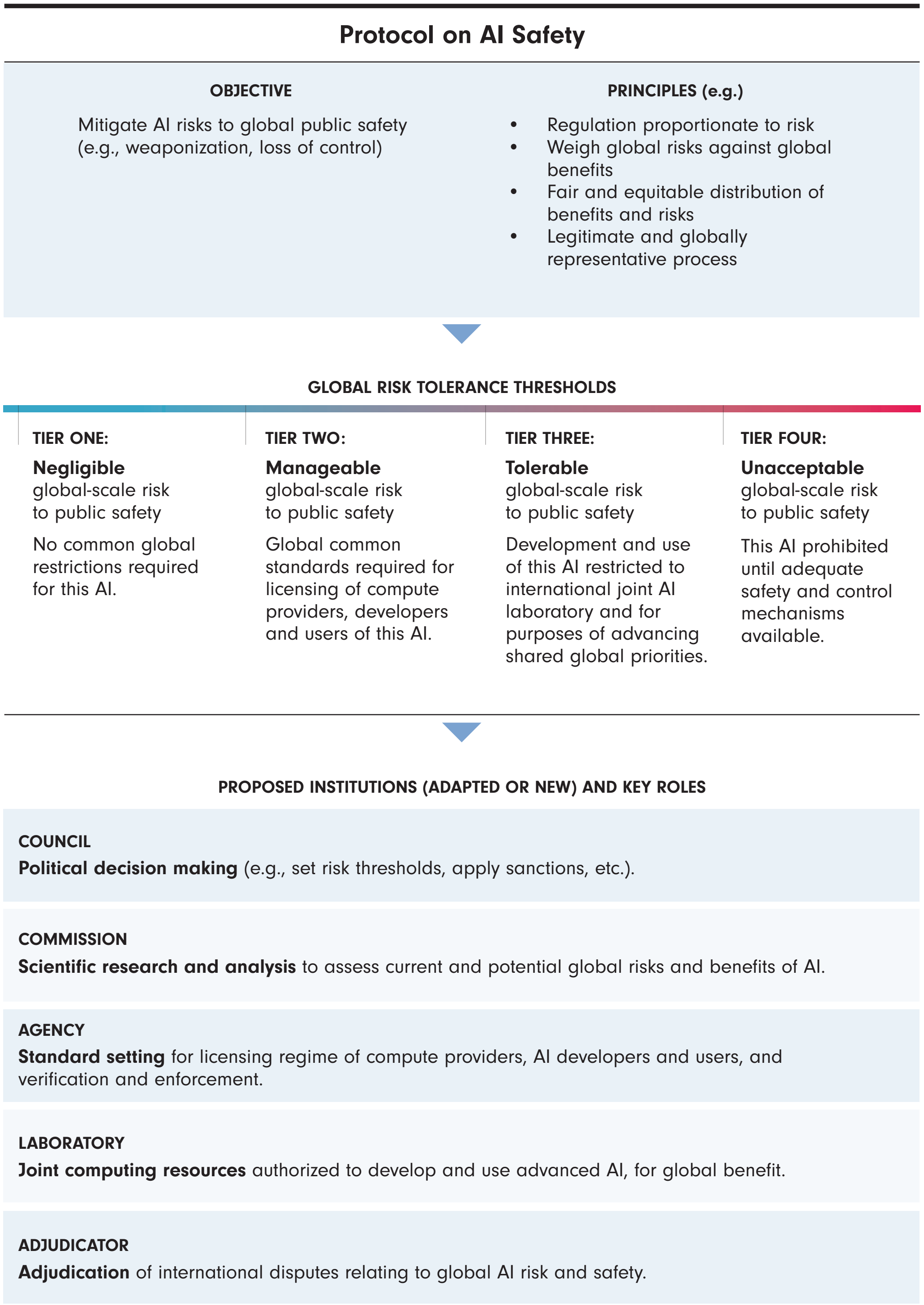
## Proposed Governance Institutions and Roles

**The following are possible institutions and roles that may be required to achieve the objectives of the Protocol on AI Safety.** Implementing the global risk threshold system described above will require several significant supporting functions, including bodies to provide scientific advice, decision-making support, technical monitoring and evaluation, and dispute and infringement adjudication. The proposed institutions could be based in or adapted from existing entities or developed new if required. The institutions that may be needed are listed below, with a more detailed description of the key functions of each body included in Appendix 2.

- **Council:** Political decision making on key issues relating to global AI risk and safety.
- **Commission:** Scientific research and analysis of state of global AI risk and safety.
- **Agency:** Design, implementation, monitoring and enforcement of shared standards and regulatory system on global AI risk and safety.
- **Laboratory:** Development and provision of AI systems for AI safety research and other global benefits. Only computing facility authorized to train and run AI systems above the maximum threshold permitted for licensed development.
- **Adjudicator:** Adjudication of international law and of disputes relating to global AI risk and safety.

Figure 5 summarizes the key components of the Protocol, including its objective and principles, four proposed tiers of action based on global risk tolerance thresholds, and the five proposed supporting institutions and their roles.

Figure 5: Protocol on Global Public Safety and Security Risks from AI



### Protocol on AI Safety

**OBJECTIVE**

Mitigate AI risks to global public safety (e.g., weaponization, loss of control)

**PRINCIPLES (e.g.)**

- Regulation proportionate to risk
- Weigh global risks against global benefits
- Fair and equitable distribution of benefits and risks
- Legitimate and globally representative process

**GLOBAL RISK TOLERANCE THRESHOLDS**

| TIER ONE: | TIER TWO: | TIER THREE: | TIER FOUR: |
|---|---|---|---|
| **Negligible** global-scale risk to public safety | **Manageable** global-scale risk to public safety | **Tolerable** global-scale risk to public safety | **Unacceptable** global-scale risk to public safety |
| No common global restrictions required for this AI. | Global common standards required for licensing of compute providers, developers and users of this AI. | Development and use of this AI restricted to international joint AI laboratory and for purposes of advancing shared global priorities. | This AI prohibited until adequate safety and control mechanisms available. |

**PROPOSED INSTITUTIONS (ADAPTED OR NEW) AND KEY ROLES**

**COUNCIL**
**Political decision making** (e.g., set risk thresholds, apply sanctions, etc.).

**COMMISSION**
**Scientific research and analysis** to assess current and potential global risks and benefits of AI.

**AGENCY**
**Standard setting** for licensing regime of compute providers, AI developers and users, and verification and enforcement.

**LABORATORY**
**Joint computing resources** authorized to develop and use advanced AI, for global benefit.

**ADJUDICATOR**
**Adjudication** of international disputes relating to global AI risk and safety.

### Illustration of the Protocol in Action

**The Protocol is designed to provide a framework for action that ensures global public safety and security under a multitude of possible future scenarios.** The following is an illustration of how the framework could function under two simplified and opposite scenarios.

- **Scenario one: AI capabilities advance more slowly than expected, accompanied by breakthroughs in successful AI control and safety techniques.** In this scenario, few if any AI systems would be capable of producing global catastrophic harms. Most AI systems could be governed at the national level, with international coordination necessary only to ensure basic common standards, such as adequate cybersecurity to prevent the hacking and dissemination of the most powerful models to malicious actors, and rules against the open release of models that could be used to cause widespread harm. Few, if any, AI models would need to be restricted to only being developed and run in the joint AI laboratory, and none would be prohibited from development entirely.

- **Scenario two: AI capabilities advance rapidly, but no reliable mechanism has yet been developed to reliably control or align AI systems.** This scenario would call for much stronger restrictions. The globally coordinated licensing regime would need to prevent anyone inside the regulated system from producing dangerous AI models and prevent anyone outside the regulated system from having access to the computing power or other inputs needed to produce dangerous AI models. Many of the most advanced AI systems would be too dangerous to be produced by individual companies or governments and would need to be restricted to only being developed and run in the joint lab. Finally, even the joint lab would hold back from training the most powerful models possible, out of concern that based on existing safeguards, such systems could possibly evade human control.

Reality could turn out to be close to either of these scenarios, or various points in between. There could also be shifts in either direction from one scenario to another, such as if sudden breakthroughs in safety research allow many more forms of AI to be considered safe or, alternatively, if sudden breakthroughs in algorithmic efficiency or falls in the cost of computing power allow many more actors (including irresponsible or malicious ones) to produce AI with dangerous capabilities. **The value of the proposed approach is that it puts in place the infrastructure needed to enable the various types of coordinated action required to address different possible scenarios as they emerge.**

# Challenges, Opportunities and Next Steps

**The proposed Framework Convention on Global AI Challenges and accompanying protocols are highly ambitious, requiring an arguably unprecedented degree of international cooperation to implement.** Achieving such cooperation may be especially challenging in the context of political polarization within societies that could undermine support for international agreement, and geopolitical tensions that could make AI development more competitive, secretive, risk-tolerant and, thus, harder to control (Danzig 2018; Clare and Ruhl

2024). Overcoming these obstacles will require a strong shared understanding of the risks and of the effective means of managing them.[7]

**Several technological hurdles will also have to be overcome to effectively implement the Framework Convention.** Enforcing risk assessment requirements for different training run thresholds may require detecting and quantifying how computing power is being used around the world, and potentially excluding some actors from leveraging it (Sastry et al. 2024). However, significant research is still required before many of the technologies such a system would require are implementation ready. More research is similarly needed to develop evaluation tools to reliably detect dangerous capabilities in new AI models, information security practices to protect critical information about advanced AI models and other technical aspects of the risk management proposals associated with the Protocol on AI Safety.

**There are multiple important governance challenges to overcome.** These include, for example, determining what degree of safety assurance should be required before the development of potentially catastrophic technologies can continue. How can such assurance be measured, and by whom? Is it even technically feasible to obtain such assurance for unprecedented technologies (Downer 2024)? There are important uncertainties about many of the specific proposals detailed above, too. This includes developing ways to incentivize states to participate in governance institutions that potentially constrain their ability to invest in and deploy new technologies, and how to verify and enforce compliance with those constraints.

**Effective international cooperation on AI may be needed very soon, given the rapid pace of advances in AI capabilities.** This creates additional challenges, as it requires a much-accelerated process for reaching international agreement on at least the core actions needed to address the most urgent and severe shared concerns. Such action may be needed before a full understanding of the risks and optimal responses is available. This also highlights the need to swiftly expand and accelerate technical and governance research to inform and improve international cooperation efforts as soon as possible.

**Despite these daunting challenges, the seriousness of the risks and the size of the potential benefits also create enormous incentives for states to cooperate to avoid disaster and secure prosperity.** Surveys indicate that the public is also wary of advanced AI and supportive of efforts to reduce risks, creating domestic political incentives for international action (Colson, n.d.; 2023). There is also precedent for international cooperation on AI and other technological risks, including evidence of shared interest among the United States and China on issues such as keeping autonomous weapons under human oversight and not integrating them into nuclear systems (Hass and Kahl 2024). Both the United States and China, as well as 26 other countries and the European Union, also signed the Bletchley Declaration (2023), which noted AI's potentially transformative effects and significant risks and the need for international collaboration to address them.

**In this context, immediate next steps to advance international cooperation on emerging global AI challenges could include:**

- engaging with diverse global stakeholders to develop and refine the proposed Framework Convention and Protocol;

- stress testing proposed measures for their likely effectiveness under challenging scenarios;

- encouraging ongoing multilateral negotiations, such as the AI Safety Summits and UN consultations, to adopt the proposed measures;

[7] The interim International Scientific Report on the Safety of Advanced AI (GOV.UK 2024a) provides an important step in building shared understanding of the risks. This, combined with additional efforts such as "track two" discussions among scientists and cooperation between national AI Safety Institutes and other bodies, could lay the foundation for the proposed international commission dedicated to this purpose. Similar levels of effort are required to build shared understanding on governance mechanisms.

- funding and expanding research and policy development efforts to fill critical knowledge gaps; and

- initiating processes to draft and adopt international agreements with the proposed measures and institutions needed to meet emerging global AI challenges.

# Conclusion

**Humanity is unprepared for the scale of global AI challenges it could soon face and lacks the mechanisms for international cooperation needed to manage such challenges effectively.** Given the pace of development in AI capabilities, a business-as-usual approach to international cooperation is unlikely to be sufficient. Existing approaches would likely prove incapable of dealing with some of the most challenging possible scenarios, such as those arising from the near-term potential to develop AI systems with vastly superhuman capabilities. That makes it prudent to design, stress test and establish effective measures in advance. This proactive approach to international cooperation on AI governance also has the advantage of limiting the risk that governments, faced with a sudden AI crisis, implement ill-considered or counterproductive responses.

**The proposed approach of a Framework Convention on Global AI Challenges, supplemented by supporting protocols on urgent and severe global-scale challenges such as risks to public safety, provides a flexible and practical way forward.** Specifically, it provides a platform for rapid universal agreement on high-level objectives and principles, combined with equally rapid concrete cooperation among key actors on the most urgent areas of shared concern. Despite this flexible framework, reaching international agreement in time will be extremely challenging. Success will require a shared focus on the common interest of humanity in safely navigating this critical juncture, combined with an ability to mobilize multiple facets of human ingenuity to solve the many technical and policy problems involved.

## Acknowledgements

The authors would like to thank the many people (from five continents) who contributed ideas and feedback to the development of this discussion paper. While their perspectives may diverge, they share a common commitment to helping humanity successfully navigate the global challenges raised by advanced AI. The authors also wish to thank the CIGI Public Affairs team for their invaluable collaboration, and the broader CIGI community for their support. Finally, the authors wish to acknowledge Paul Samson, president of CIGI, for astute and insightful guidance, as well as substantive contributions to the discussion paper.

## Appendix 1: Descriptions of AI Risks in Figure 3

| Risk | Description |
|---|---|
| Loss of job | AI is expected to cause economic dislocation as it transforms the economy (Georgieva 2024). This could significantly disrupt the lives of affected families and individuals. |
| Medical diagnostics failures | AI could revolutionize medical diagnostics, but overreliance on AI poses risks of misdiagnosis and harmful treatment choices (Macmillan 2024). |
| Autonomous vehicle fatalities | Though AI-enabled services such as self-driving cars may ultimately be safer, they will likely experience problems as they are deployed in the real world (Muzahid et al. 2023). |
| Privacy breaches | AI's reliance on vast amounts of data increases the risk of privacy breaches, potentially affecting millions (Passeri 2023). |
| Robotics malfunctions | AI and robotics technologies will allow autonomous systems to take on many more roles in society, such as in logistics and manufacturing. But especially initially, these systems could malfunction in various dangerous ways when deployed in the real world (Schneier and Ottenheimer 2023). |
| Behavioural manipulations | AI may be used to generate persuasive, highly tailored propaganda that could make it easier to manipulate people on a large scale (Burtell and Woodside 2023). |
| Algorithmic bias | As AI systems gain influence over societal decisions, biased decision making threatens to disadvantage specific groups. AI bias impacts critical areas such as hiring, policing, incarceration, marketing and insurance (Mehrabi et al. 2021). |
| Mass job loss | AI will boost productivity, but also may lead to economic upheaval as prices fall and jobs shift (Bughin et al. 2018). These disruptions will fall disproportionately on certain groups and sectors. |
| Cyberattacks on critical infrastructure | According to the United Kingdom's National Cyber Security Centre (2024), AI will "almost certainly" increase the volume and severity of cyberattacks in coming years. These attacks pose a threat to individuals, financial institutions, public services, corporate information and government infrastructure. |
| Hyper-surveillance | Cheap and effective AI surveillance could threaten civil liberties. AI surveillance tools are already in use in at least 75 countries (Feldstein 2019). |
| Harm to democratic institutions | AI systems could be used to generate individually tailored, persuasive propaganda, threatening democratic systems by degrading the information environment (Matz et al. 2024). |
| Automated warfare | AI is expected to have many military uses, from improved sensors and targeting to lethal autonomous weapon systems. AI systems could increase the speed at which war is fought, as they collect and process data much faster than human operators. Militaries may be incentivized to use AI systems whenever possible for fear of being outmanoeuvred by their adversaries. This has led some researchers to raise worries about the possibility of hyperwar, a war directed by AI systems that moves faster than any person can understand or control (Scharre 2023). Such a war may escalate extremely quickly, becoming a flash war (Johnson 2022). |

| Risk | Description |
|---|---|
| AI-enabled totalitarianism | AI systems excel at autonomously collecting and processing enormous amounts of data. All governments will have to manage risks from enhanced surveillance and control, but in autocratic regimes the prospect of an "AI-tocracy" is particularly concerning (Beraja 2023). In the worst case, oppressive AI-powered regimes may prove impossible to dislodge (Caplan 2008). |
| Widening global inequality | As AI systems become more powerful and economically useful, they could allow the companies or individuals who control them to amass power and unduly influence political decision making through regulatory capture and other mechanisms (Nolan 2023). AI may also benefit high-income workers or owners of capital the most, exacerbating global inequality (Georgieva 2024). |
| Weaponization | AI tools could make it easier to access information and synthesize dangerous pathogens (Batalis 2023). This could increase the threat posed by states, groups or individuals in developing and deploying bioweapons. Many AI applications in biotechnology and other areas will be dual use (Urbina et al. 2022). |
| Uncontrolled superintelligence | It may be difficult to train AI systems that do exactly what users intend, especially when they are performing complicated tasks in the real world. Many researchers are concerned that competitive pressures will lead AI developers to release models that end up taking actions that harm people (Ji 2024).<br>For example, an AI could be using tricks or shortcuts to solve problems in its training data that have disastrous effects in the real world (OpenAI 2023). Or the goals the system learned in training might not work well in the full range of real-world scenarios (Shah et al. 2022). Or, perhaps worst of all, a superintelligent AI system could "realize" that — whatever its goals — amassing power and disabling opportunities to turn it off would be helpful (Carlsmith 2022). Such a system may appear cooperative at first but take over global systems once it has the chance (Hubinger et al. 2024).<br>Such a system could be powerful enough to threaten all of humanity (Piper 2020). |

## Appendix 2: Summary of Proposed Institutions, Roles and Functions

This table summarizes the possible institutions, roles and key functions that may be required to achieve the objectives of the Protocol on AI Safety. The proposed institutions could be adapted from existing entities or developed new if required. These institutions could also serve functions of other protocols under the proposed Framework Convention on Global AI Challenges.

| Proposed Institution | Role | Possible Key Functions |
|---|---|---|
| **Council** | Political decision making on key issues relating to global AI risk and safety. | • **Facilitate effective and legitimate decision making** on key collective decisions to ensure global AI safety, including:<br>– amendments to protocol;<br>– risk threshold tiers;<br>– priorities for joint AI development; and<br>– enforcement actions.<br>• Consider recommendations of commission, agency and laboratory.<br>• Conduct deliberative assemblies. |
| **Commission** | Scientific research and analysis of state of global AI risk and safety. | • **Assess risk levels of various types of AI models and systems** based on best available empirical evidence (e.g., capabilities evaluations) as well as proxies (e.g., compute usage, model parameters) for measuring them.<br>• Update assessed risk levels of various types of AI models on an ongoing basis based on technological developments, safety improvements and improved available data and research.<br>• Provide technical advice and recommendations to the council and the agency. |

| Proposed Institution | Role | Possible Key Functions |
|---|---|---|
| **Agency** | Design, implementation, monitoring and enforcement of shared standards and regulatory system on global AI risk and safety. | • Develop common standards for licensing of compute providers, AI developers and AI users, as required for ensuring safety from global AI risk.<br>• Develop international and recommended national legal framework for liability of compute providers, AI developers and AI users, as required for ensuring safety from global AI risk.<br>• **Accredit** national or other AI safety organizations responsible for licensing and auditing compute providers, AI model developers and AI model users related to global catastrophic risk from AI.<br>• Provide direct services (e.g., licensing and audit) in countries lacking access to a national or other accredited body.<br>• Conduct **monitoring** of state party compliance with the provisions of the agreement, including through jurisdiction-level and firm-level monitoring.<br>• Set guidelines on the appropriate and justified **enforcement** of the provisions of the agreement.<br>– Recommend to member states the appropriate enforcement mechanisms in response to a given instance of non-compliance. |
| **Laboratory** | Provision of compute resources and development of AI for AI safety and other global benefits. | • Manage and run the only **computing facility** that is authorized to train AI systems above the maximum threshold permitted for licensed development.<br>• **Develop and run AI systems** above the maximum threshold permitted for licensed development, where authorized by the council on recommendation of the agency, and when justified by the potential global benefits.<br>• **Conduct AI safety research** on behalf of the global community. |
| **Adjudicator** | Adjudication of international law and of disputes relating to global AI risk and safety. | • **Adjudicate international law related to global AI risk and safety.**<br>– Adjudicate based on existing international law and on new legal frameworks adopted by the council.<br>– Adjudicate disputes between state parties.<br>– Adjudicate cases in jurisdictions that lack adequate legal frameworks (e.g., liability regimes).<br>• Provide findings and recommend action (e.g., remedies, penalties) to council. |

# Works Cited

Advisory Body on Artificial Intelligence. 2023. *Interim Report: Governing AI for Humanity.* December. New York, NY: United Nations. www.un.org/sites/un2.un.org/files/un_ai_advisory_body_governing_ai_for_humanity_interim_report.pdf.

African Union. 2024. "African Ministers Adopt Landmark Continental Artificial Intelligence Strategy, African Digital Compact to drive Africa's Development and Inclusive Growth." Press release, June 17. https://au.int/en/pressreleases/20240617/african-ministers-adopt-landmark-continental-artificial-intelligence-strategy.

Anderljung, Markus, Joslyn Barnhart, Anton Korinek, Jade Leung, Cullen O'Keefe, Jess Whittlestone, Shahar Avin et al. 2023. "Frontier AI Regulation: Managing Emerging Risks to Public Safety." *arXiv*, November 7. https://doi.org/10.48550/arXiv.2307.03718.

Batalis, Steph. 2023. "AI and Biorisk: An Explainer." Center for Security and Emerging Technology, December. https://cset.georgetown.edu/publication/ai-and-biorisk-an-explainer/

Bengio, Yoshua, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Trevor Darrell, Yuval Noah Harari et al. 2024. "Managing extreme AI risks amid rapid progress." *Science* 384 (6698): 842–45. https://doi.org/10.1126/science.adn0117.

Beraja, Martin, Andrew Kao, David Y. Yang and Noam Yuchtman. 2023. "AI-tocracy." *The Quarterly Journal of Economics* 138 (3): 1349–1402. https://doi.org/10.1093/qje/qjad012.

Bowman, Samuel R. 2023. "Eight Things to Know about Large Language Models." *arXiv*, April 2. http://arxiv.org/abs/2304.00612.

Bughin, Jacques, Jeongmin Seong, James Manyika, Michael Chui and Raoul Joshi. 2018. "Notes from the AI frontier: Modeling the impact of AI on the world economy." McKinsey Global Institute, September 4. www.mckinsey.com/featured-insights/artificial-intelligence/notes-from-the-ai-frontier-modeling-the-impact-of-ai-on-the-world-economy.

Burtell, Matthew and Thomas Woodside. 2023. "Artificial Influence: An Analysis Of AI-Driven Persuasion."*arXiv*, March 15. https://doi.org/10.48550/arXiv.2303.08721

Caplan, Bryan. 2008. "The totalitarian threat." In *Global Catastrophic Risks*, edited by Nick Bostrom and Milan M. Ćirković, 504–30. Oxford, UK: Oxford University Press. https://doi.org/10.1093/oso/9780198570509.003.0029.

Carlsmith, Joe. 2022. "Is Power-Seeking AI an Existential Risk?" *arXiv*, June 16. https://doi.org/10.48550/arXiv.2206.13353.

Cihon, Peter. 2019. *Standards for AI Governance: International Standards to Enable Global Coordination in AI Research & Development*. Future of Humanity Institute Technical Report. April. www.fhi.ox.ac.uk/wp-content/uploads/Standards_-FHI-Technical-Report.pdf.

Clare, Stephen and Christian Ruhl. 2024. *Great Power Competition and Transformative Technologies.* Founders Pledge. January. www.founderspledge.com/research/great-power-competition-and-transformative-technologies-report.

Colson, Daniel. n.d. "New Poll Finds Preventing Catastrophic Outcomes is the Top AI Policy Objective for Americans, Majority Support Regulation of Deepfakes, and Ban on AI-Written News Articles." Artificial Intelligence Policy Institute. https://theaipi.org/poll-biden-ai-executive-order-10-30-2/.

———. 2023. "Overwhelming Majority of Voters Believe Tech Companies Should be Liable for Harm Caused by AI Models, Favor Reducing AI Proliferation and Law Requiring Political Ad Disclose Use of AI." Artificial Intelligence Policy Institute, September 19. https://theaipi.org/poll-shows-voters-oppose-open-sourcing-ai-models-support-regulatory-representation-on-boards-and-say-ai-risks-outweigh-benefits-2/.

Council of Europe. 2024. "Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law." Council of Europe Treaty Series, May 10. CETS 225 - Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law. www.coe.int/en/web/artificial-intelligence/the-framework-convention-on-artificial-intelligence.

Critch, Andrew and Stuart Russell. 2023. "TASRA: A Taxonomy and Analysis of Societal-Scale Risks from AI." *arXiv*, June 14. http://arxiv.org/abs/2306.06924.

Danzig, Richard. 2018. *Technology Roulette: Managing Loss of Control as Many Militaries Pursue Technological Superiority*. Washington, DC: Center for a New American Security.

Downer, John. 2024. *Rational Accidents: Reckoning with Catastrophic Technologies*. Cambridge, MA: MIT Press.

Everitt, Tom, Gary Lea and Marcus Hutter. 2018. "AGI safety literature review." *arXiv*, May 21. https://doi.org/10.48550/arXiv.1805.01109.

Fang, Richard, Rohan Bindu, Akul Gupta, Qiusi Zhan and Daniel Kang. 2024. "LLM Agents can Autonomously Hack Websites." *arXiv*, February 16. https://doi.org/10.48550/arXiv.2402.06664.

Feldstein, Steven. 2019. "The Global Expansion of AI Surveillance." Carnegie Endowment for International Peace, September 17. https://carnegieendowment.org/research/2019/09/the-global-expansion-of-ai-surveillance?lang=en.

G7. 2023. "Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems." G7 2023 Hiroshima Summit, October 30. www.mofa.go.jp/files/100573471.pdf.

Garfinkel, Ben and Allan Dafoe. 2021. "How does the offense-defense balance scale?" In *Emerging Technologies and International Stability*, 1st ed., edited by Todd S. Sechser, Neil Narang and Caitlin Talmadge, 247–74. Abingdon, UK: Routledge.

Georgieva, Kristalina. 2024. "AI Will Transform the Global Economy. Let's Make Sure It Benefits Humanity." *IMF Blog*, January 14. www.imf.org/en/Blogs/Articles/2024/01/14/ai-will-transform-the-global-economy-lets-make-sure-it-benefits-humanity.

Giattino, Charlie, Edouard Mathieu, Veronika Samborska and Max Roser. 2023. "Artificial Intelligence." Our World in Data. https://ourworldindata.org/artificial-intelligence.

GOV.UK. 2023. "The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023." www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchleydeclaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023.

———. 2024a. *International Scientific Report on the Safety of Advanced AI: Interim Report*. May 17. www.gov.uk/government/publications/international-scientific-report-on-the-safety-of-advanced-ai.

———. 2024b. "Seoul Declaration for safe, innovative and inclusive AI by participants attending the Leaders' Session: AI Seoul Summit, 21 May 2024." Department for Science, Innovation & Technology Policy Paper, May 21. www.gov.uk/government/publications/seoul-declaration-for-safe-innovative-and-inclusive-ai-ai-seoul-summit-2024/seoul-declaration-for-safe-innovative-and-inclusive-ai-by-participants-attending-the-leaders-session-ai-seoul-summit-21-may-2024.

———. 2024c. "Seoul Statement of Intent toward International Cooperation on AI Safety Science, AI Seoul Summit 2024 (Annex)." Department for Science, Innovation & Technology Policy Paper, May 21. www.gov.uk/government/publications/seoul-declaration-for-safe-innovative-and-inclusive-ai-ai-seoul-summit-2024/seoul-statement-of-intent-toward-international-cooperation-on-ai-safety-science-ai-seoul-summit-2024-annex.

Hass, Ryan and Colin Kahl. 2024. "Laying the groundwork for US-China AI dialogue." Brookings Institution, April 5. www.brookings.edu/articles/laying-the-groundwork-for-us-china-ai-dialogue/.

Heim, Lennart, Tim Fist, Janet Egan, Sihao Huang, Stephen Zekany, Robert Trager, Michael A. Osborne and Noa Zilberman. 2024. "Governing Through the Cloud: The Intermediary Role of Compute Providers in AI Regulation." *arXiv*, March 26. https://doi.org/10.48550/arXiv.2403.08501.

Hill, Michael. 2023. "Generative AI phishing fears realized as model develops 'highly convincing' emails in 5 minutes." CSO, October 24. www.csoonline.com/article/656698/generative-ai-phishing-fears-realized-as-model-develops-highly-convincing-emails-in-5-minutes.html.

Horowitz, Michael C. 2018. "Artificial Intelligence, International Competition, and the Balance of Power." *Texas National Security Review* 1 (3): 36–57. https://doi.org/10.15781/T2639KP49.

Hubinger, Evan, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham et al. 2024. "Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training." *arXiv*, January 17. https://doi.org/10.48550/arXiv.2401.05566.

International Nuclear Safety Advisory Group. 1999. *Basic Safety Principles for Nuclear Power Plants — 75-INSAG-3 Rev. 1*. Vienna, Austria: International Atomic Energy Agency.

International Telecommunication Union. 2023. *Measuring digital development: Facts and Figures 2023*. www.itu.int/itu-d/reports/statistics/facts-figures-2023/.

Ji, Jiaming, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan et al. 2024. "AI Alignment; A Comprehensive Survey." *arXiv*, May 1. https://doi.org/10.48550/arXiv.2310.19852.

Johnson, James. 2022. "AI, Autonomy, and the Risk of Nuclear War." War on the Rocks, July 29. https://warontherocks.com/2022/07/ai-autonomy-and-the-risk-of-nuclear-war/.

Kinniment, Megan, Lucas Jun Koba Sato, Haoxing Du, Brian Goodrich, Max Hasin, Lawrence Chan, Luke Harold Miles et al. 2024. "Evaluating Language-Model Agents on Realistic Autonomous Tasks." *arXiv*, January 4. https://doi.org/10.48550/arXiv.2312.11671.

Koessler, Leonie, Jonas Schuett and Markus Anderljung. 2024. "Risk thresholds for frontier AI." *arXiv*, June 20. https://arxiv.org/abs/2406.14713.

Maas, Matthijs M. 2023. "Concepts in Advanced AI Governance: A Literature Review of Key Terms and Definitions." *AI Foundations Report 3*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4612473.

Macmillan, Carrie. 2024. "Generative AI for Health Information: A Guide to Safe Use." Yale Medicine, January 8. www.yalemedicine.org/news/generative-ai-artificial-intelligence-for-health-info.

Matz, Sandra, Jake Teeny, Sumer S. Vaid, Gabriella M. Harari and Moran Cerf. 2024. "The potential of generative AI for personalized persuasion at scale." *Scientific Reports* 14 (1): 4692. https://doi.org/10.1038/s41598-024-53755-0.

Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman and Aram Galstyan. 2021. "A Survey on Bias and Fairness in Machine Learning." *ACM Computing Surveys* 54 (6): 1–35. https://doi.org/10.1145/3457607.

Ministry of Foreign Affairs, People's Republic of China. 2023. "Global AI Governance Initiative." October 30. www.mfa.gov.cn/eng/wjdt_665385/2649_665393/202310/t20231020_11164834.html.

Morris, Meredith Ringel, Jascha Sohl-Dickstein, Noah Fiedel, Tris Warkentin, Allan Dafoe, Aleksandra Faust, Clement Farabet and Shane Legg. 2024. "Position: Levels of AGI for Operationalizing Progress on the Path to AGI." *Proceedings of the 41st International Conference on Machine Learning*. Vienna, Austria. https://arxiv.org/pdf/2311.02462.

Muzahid, Abu Jafar Md, Syafiq Fauzi Kamarulzaman, Md Arafatur Rahman, Saydul Akbar Murad, Md Abdus Samad Kamal and Ali H. Alenezi. 2023. "Multiple vehicle cooperation and collision avoidance in automated vehicles: Survey and an AI-enabled conceptual framework." *Scientific Reports* 13 (1): 603. www.nature.com/articles/s41598-022-27026-9.

National Cyber Security Centre. 2024. "The near-term impact of AI on the cyber threat." UK Government, January 24. www.ncsc.gov.uk/report/impact-of-ai-on-cyber-threat.

Nolan, Beatrice. 2023. "Don't let Big Tech write the AI rules, warns AI godfather." *Business Insider*, November 4. www.businessinsider.com/big-tech-controlling-ai-sector-concerns-ai-godfather-yoshua-bengio-2023-11.

O'Keefe, Cullen, Peter Cihon, Ben Garfinkel, Carrick Flynn, Jade Leung and Allan Dafoe. 2020. *The Windfall Clause: Distributing the Benefits of AI for the Common Good*. Future of Humanity Institute, University of Oxford. www.fhi.ox.ac.uk/wp-content/uploads/Windfall-Clause-Report.pdf.

OpenAI. 2023. "GPT-4 System Card." March 23. https://cdn.openai.com/papers/gpt-4-system-card.pdf.

Passeri, Paolo. 2023. "The Risk of Accidental Data Exposure by Generative AI is Growing." *Infosecurity Magazine* (blog), August 16. www.infosecurity-magazine.com/blogs/accidental-data-exposure-gen-ai/.

Peterson, Dahlia and Samantha Hoffman. 2022. "Geopolitical implications of AI and digital surveillance adoption." Brookings Institution, June. www.brookings.edu/articles/geopolitical-implications-of-ai-and-digital-surveillance-adoption/.

Piper, Kelsey. 2020. "The case for taking AI seriously as a threat to humanity." *Vox*, October 15. www.vox.com/future-perfect/2018/12/21/18126576/ai-artificial-intelligence-machine-learning-safety-alignment.

Sandbrink, Jonas B. 2023. "Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools." *arXiv*, December 23. https://doi.org/10.48550/arXiv.2306.13952.

Sastry, Girish, Lennart Heim, Haydn Belfield, Markus Anderljung, Miles Brundage, Julian Hazell, Cullen O'Keefe et al. 2024. "Computing Power and the Governance of Artificial Intelligence." *arXiv*, February 13. https://doi.org/10.48550/arXiv.2402.08797.

Scharre, Paul. 2023. *Four battlegrounds: power in the age of artificial intelligence*. W. W. Norton & Company.

Schneier, Bruce and Davi Ottenheimer. 2023. "Robots Are Already Killing People." *The Atlantic*, September 6. www.theatlantic.com/technology/archive/2023/09/robot-safety-standards-regulation-human-fatalities/675231/.

Sebo, Jeff and Robert Long. 2023. "Moral consideration for AI systems by 2030." *AI and Ethics*. https://doi.org/10.1007/s43681-023-00379-1.

Sevilla, Jaime and Edu Roldán. 2024. "Training Compute of Frontier AI Models Grows by 4–5x Per Year." Epoch AI, May 28. https://epochai.org/blog/training-compute-of-frontier-ai-models-grows-by-4-5x-per-year.

Shah, Rohin, Vikrant Varma, Ramana Kumar, Mary Phuong, Victoria Krakovna, Jonathan Uesato and Zac Kenton. 2022. "Goal Misgeneralization: Why Correct Specifications Aren't Enough For Correct Goals." *arXiv*, November 2. https://doi.org/10.48550/arXiv.2210.01790.

Shevlane, Toby, Sebastian Farquhar, Ben Garfinkel, Mary Phuong, Jess Whittlestone, Jade Leung, Daniel Kokotajlo et al. 2023. "Model evaluation for extreme risks." *arXiv*, September 22. http://arxiv.org/abs/2305.15324.

Trager, Robert F., Ben Harack, Anka Reuel, Allison Carnegie, Lennart Heim, Lewis Ho, Sarah Kreps et al. 2023. "International Governance of Civilian AI: A Jurisdictional Certification Approach." *arXiv*, September 11. https://doi.org/10.48550/arXiv.2308.15514.

United Nations. 2021. *Our Common Agenda: Report of the Secretary-General*. New York, NY: United Nations. www.un.org/en/content/common-agenda-report/assets/pdf/Common_Agenda_Report_English.pdf.

United Nations General Assembly. 2024. "Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development." A/78/L.49. https://documents.un.org/doc/undoc/ltd/n24/065/92/pdf/n2406592.pdf?token=C67TWfiAvlnNGhnG7H&fe=true.

Urbina, Fabio, Filippa Lentzos, Cédric Invernizzi and Sean Ekins. 2022. "Dual use of artificial-intelligence-powered drug discovery." *Nature Machine Intelligence* 4 (3): 189–91. https://doi.org/10.1038/s42256-022-00465-9.

## About CIGI

The Centre for International Governance Innovation (CIGI) is an independent, non-partisan think tank whose peer-reviewed research and trusted analysis influence policy makers to innovate. Our global network of multidisciplinary researchers and strategic partnerships provide policy solutions for the digital era with one goal: to improve people's lives everywhere. Headquartered in Waterloo, Canada, CIGI has received support from the Government of Canada, the Government of Ontario and founder Jim Balsillie.

## À propos du CIGI

Le Centre pour l'innovation dans la gouvernance internationale (CIGI) est un groupe de réflexion indépendant et non partisan dont les recherches évaluées par des pairs et les analyses fiables incitent les décideurs à innover. Grâce à son réseau mondial de chercheurs pluridisciplinaires et de partenariats stratégiques, le CIGI offre des solutions politiques adaptées à l'ère numérique dans le seul but d'améliorer la vie des gens du monde entier. Le CIGI, dont le siège se trouve à Waterloo, au Canada, bénéficie du soutien du gouvernement du Canada, du gouvernement de l'Ontario et de son fondateur, Jim Balsillie.

Centre for International
Governance Innovation

67 Erb Street West
Waterloo, ON, Canada N2L 6C2
www.cigionline.org