United Nations University
Centre for Policy Research

WORKINGPAPER

September 2023

Towards a UN Role in Governing Foundation Artificial Intelligence Models

Jason Hausenloy, Claire Dennis

Towards a UN Role in Governing Foundation Artificial Intelligence Models

Jason Hausenloy
United Nations University

Claire Dennis
Princeton University

Executive Summary

"Alarm bells over the latest form of artificial intelligence – generative AI – are deafening. And they are loudest from the developers who designed it…We must take those warnings seriously."

- António Guterres, UN Secretary-General, June 2023[1]

Propelled by rapid progress in Artificial Intelligence (AI), scientists and policymakers are increasingly advocating for international governance to steer this transformative technology toward the global public good. In particular, these calls focus on 'foundation models,' which, as the name suggests, are designed to be the foundation for a wide variety of tasks and applications. Today's foundation models most notably include large language models like GPT-4 and image generators like DALLE-2. Given the exponential improvement in the capabilities of these models within the last few years – indeed, even in just the last few months – there is a growing sense of alarm over the harms and potential risks, even existential threats, that their proliferation poses if left unchecked.

Within the UN system, the Secretary-General and the Envoy on Technology's Office, among many other offices, are leading a range of initiatives, including the recently announced Multistakeholder Advisory Body on AI (Advisory Body), which could inform the eventual creation of a multilateral AI governance institution. **This report aims to contribute to the UN's consideration of AI governance strategies in the short-term and frame the agenda of the Multistakeholder Advisory Body on AI.**

This report is structured into three distinct parts, from which we extract the main arguments:

PART I: Foundation AI models and their unique challenges should be the central focus of the UN's AI governance efforts.

- The rapid development of advanced AI systems known as foundation models poses governance challenges and risks if left unregulated. This section examines the key issues in governing foundation AI during the phases of development, deployment, and oversight.
- Foundation models can progress unexpectedly fast and lead to societal risks from misuse, proliferation, and automation. However, their opacity and evaluation limitations impede effective governance.

[1] "Secretary-General Urges Broad Engagement from All Stakeholders towards United Nations Code of Conduct for Information Integrity on Digital Platforms | UN Press." Un.org, 12 June 2023, press.un.org/en/2023/sgsm21832.doc.htm.

- Safety expertise and compute resources are highly concentrated within a few private sector companies, mostly in the US. This constrains international regulatory capabilities, especially for the UN.

PART II: No existing institutional model can be 'copy-pasted' to respond to the risks from frontier AI.

- Current proposals for international institutions highlight important governance mechanisms, but have significant shortcomings in enforceability, agility, and applicability.
- International AI governance cannot be achieved by copy-pasting existing models, but rather by using these historical examples to employ a multi-pronged approach.

PART III: Any UN international institution charged with AI governance responsibilities should focus on norm and consensus-building, rather than a technical regulatory mechanism.

- Given the UN's limited technical oversight capabilities in Foundation AI, the Advisory Body should propose an international regime which amplifies the voices of less powerful actors, advocates for equitable distribution of benefits, and builds consensus around universal norms within AI.
- The rapid development of AI poses risks of a "tragedy of the commons" if left ungoverned, threatening humanity's collective capacity to adapt to rapid technological change. The UN is uniquely positioned as a global authority to respond to this dilemma.

Our specific recommendations for the Multistakeholder Advisory Body on AI to build the UN's capacity for foundation model governance are:

1. Focus efforts on maintaining human control to mitigate harms and extreme risks, starting at the research and development stage of frontier AI.
2. Drive international convergence around best practices in frontier AI governance such as risk assessments, model evaluations, and hardware controls.
3. Convene inclusive multinational consultations on the acceptable global risks from the development of AI, and advocate for the halting of dangerous research.
4. Engage leading AI companies to build technical understanding of the dangers frontier AI development poses, and submit companies' plans for AI safety and alignment to international scrutiny.
5. Support and advise the development of effective regulation, particularly in countries that lack capacity to ensure technological development proceeds at a manageable pace for societies worldwide.
6. Build safe spaces for high-trust skill and knowledge transfer between leading AI developers and regulators to boost technical understanding within governing bodies.

7. Boost multi-stakeholder fora for the public accountability of AI developers and regulators to ensure increasing adherence to international norms and guidelines.
8. Support low- and middle-income countries in engaging with governance processes to fairly shape representation and develop effective benefit-sharing mechanisms.

The UN has an opportunity to shape the development of AI for the global good. This report proposes a pragmatic approach that plays to the UN's strengths in building international norms and consensus. Rather than rushing to create enforceable international regulations, this proposed strategy allows the UN to harness its convening power and moral authority while buying time for technical regulatory capacity to grow – gradually laying the groundwork for the emergence of an effective international regime complex for AI.

Part I:
Understanding the Unique Challenges of Artificial Intelligence Governance

Overview Part I:

I. Abstract:

The rapid development of advanced artificial intelligence (AI) systems known as foundation models poses governance challenges and risks if left unregulated. This section examines the key issues in governing foundation AI during the phases of development, deployment, and oversight.

Key findings:

- Foundation models can progress unexpectedly fast and lead to societal risks from misuse, proliferation, and automation. However, their opacity and evaluation limitations impede effective governance.
- Safety expertise and compute resources are highly concentrated within a few private sector companies, mostly in the US. This constrains international regulatory capabilities, especially for the UN.

Key Definitions:

Artificial intelligence (AI): "A set of techniques aimed at approximating some aspect of biological or human intelligence in machines." (AI Governance: A Research Agenda)

Foundation models: "A foundation model is a model trained on broad data at scale in order to be generally useful across tasks." (Bommasani et al., 2021)

Frontier AI: "Frontier AI refers to the cutting edge of artificial intelligence capabilities in a given era." (Bridging AI's Economic Impact Gaps, 2022)

Generative AI: "Generative artificial intelligence (AI) refers to AI systems based on machine learning techniques that are capable of 'generating' various forms of data and content." (Anthropic Blog)

Large language models (LLMs): Large language models generally refer to language models that have hundreds of millions (and at the cutting edge, hundreds of billions) of parameters, which are pretrained using billions of words of text and use a transformer neural network architecture.[2]

Compute: The computational resources required for artificial intelligence systems to perform tasks, such as processing data, training machine learning models, and making predictions.

(Global) Catastrophic Risk: The probability of events "that result in over 10 million fatalities, or greater than $10 trillion in damages, essentially the damage must be extensive and on a global scale." (Global Assessment report on Disaster Risk Reduction 2022)

Existential Risk: The probability of human extinction or the irreversible end of development over a given timeframe. (UNDRR Thematic Study 2023)

[2] 'GPT' in ChatGPT stands for *generative pre-trained transformer* -- a type of architecture for LLMs.

I. What are Foundation Models?

Demystifying Definitions

There are various terms used today to describe leading AI systems, including 'foundation models,' 'frontier models,' and 'generative AI.' These terms are often confused and merit clear explanation. "Foundation models,"[3] are AI models trained on large datasets that can be adapted to a wide range of downstream tasks. Foundation models underlie the vast majority of AI products being developed today, such as ChatGPT, a chatbot, and image generators like Stable Diffusion.[4] 'Generative AI' refers to AI systems that can be used to create new content, including audio, code, images, text, and videos. Not all generative AI models are foundation models; for example, Amazon's Alexa is a generative AI model, but not a foundation model.

'Frontier AI,' while lacking a consistent definition, in practice generally refers to the most cutting-edge, powerful models of a given era.[5] Frontier AI systems today are foundation models, as these are the models that currently exhibit the newest and most advanced capabilities.[6] As technologies develop, however, today's frontier models will eventually be replaced by even more powerful ones, which may be different from foundation models. There is no agreed-upon way of measuring whether a model is 'frontier' or not, though currently, the computational resources needed to train the model is a proxy that is sometimes used – as it is measurable, and generally, larger systems are more powerful. However, this correlation could diminish as future algorithms become more efficient and require fewer and fewer compute resources.

In this report, we focus on **foundation models** as the most concrete term to encapsulate today's advanced AI systems, noting that policymakers will have to account for the breakneck speed of development and develop the internal technical expertise to continuously update their understanding of terms and concepts and adapt regulation accordingly.

Speed, Size, and Progress

Foundation models are evolving extremely rapidly. It is not difficult to imagine a near-term scenario in which it becomes difficult to discern whether one is interacting with a human or a machine. Capabilities have jumped from an AI beating a human in chess in 1997,[7] to AIs today

[3] https://arxiv.org/abs/2108.07258
[4] https://media.un.org/en/asset/k14/k14ar7aqzw
[5] https://www.adalovelaceinstitute.org/resource/foundation-models-explainer/#:~:text='Frontier%20models'%20are%20a%20type,by%20industry%2C%20policymakers%20and%20regulators.
[6] In reality, Frontier AI systems are trained with a number of techniques, not just prediction from large datasets, but also, for example, human feedback and self-play (the most prominent example of which combining all three is the planned release of Gemini by Google Deepmind, which uses the AlphaGo algorithms to improve the performance of large language models, text-based foundation models.
[7] https://www.ibm.com/ibm/history/ibm100/us/en/icons/deepblue/

that are able to simulate autonomous pilots, stabilize plasma in fusion reactors, and design semiconductors.[8]

Foundation models are also extremely large and expensive. To illustrate the scale of these systems, GPT-4, the latest foundation model trained by OpenAI, required hundreds of millions of dollars and enormous computing power to train. This foundation model was trained to predict the next word in a given sentence, allowing it to learn to code and power chatbot services like ChatGPT. However, when combined with extensions like AutoGPT or 'Plug-ins,' the model can develop detailed plans and execute actions autonomously. Fortunately, these plans are not yet strong enough to pose a danger, but that could quickly change.[9]

Benefits and Risks

Advances in frontier AI are a double-edged sword. In a recent UN Security Council meeting, Secretary-General Guterres remarked that, "Generative AI has enormous potential for good and evil at scale."[10]

AI technology promises transformative social, political, and economic benefits for all nations. Researchers across disciplines are using AI for data analysis and discovery problems, promising breakthroughs in scientific discovery, software development, and healthcare. Language models are estimated to be writing 3 per cent of code at Google.[11] As José Gonçlaves, the Deputy Minister for Foreign Affairs of Mozambique noted, AI technologies could help eradicate disease, combat climate change, and customize mediation efforts.[12] The leading AI developers , OpenAI, Google Deepmind, and Anthropic, are even more ambitious.[13] They are not only making chatbots, but aim to create "highly autonomous systems that outperform humans at most economically valuable work."[14]

However, without regulatory intervention, unconstrained development of frontier AI systems could concentrate power in the hands of a few private actors, leading to the mass automation of a variety of jobs that could result in unemployment, and in the worst case scenario, create catastrophic and existential risks.[15]

---

[8] Jack Clark, in his remarks to the UNSC said: "AI systems…can do things as varied as: develop autonomous fighter pilots that can beat humans in military simulations, stabilize the plasma in fusion reactors, and even design the layout of next-generation semiconductors."
[9] https://www.economist.com/by-invitation/2023/07/21/one-of-the-godfathers-of-ai-airs-his-concerns
[10] https://press.un.org/en/2023/sgsm21880.doc.htm
[11] https://ai.googleblog.com/2022/07/ml-enhanced-code-completion-improves.html
[12] https://press.un.org/en/2023/sc15359.doc.htm
[13] https://openai.com/charter, https://visualisingai.deepmind.com/theme/artificial-general-intelligence, https://www.anthropic.com/index/core-views-on-ai-safety,
[14] https://openai.com/charter
[15] In his address to the UN Security Council, Jack Clark likened AI to a type of human labour – "one that can be bought and sold at the speed of a computer, and one which is getting cheaper and more capable over time. And, as I have just described, this is a form of labour that is being developed by one narrow class of actors – companies. We should be clear eyed about the immense political leverage this affords –

Governing foundation models at an international level seems necessary for their safe development and inclusive use. However, these models have a number of features that make them difficult to regulate on an international level in the short-term, that any effective governance scheme must address.

II. Challenges in Foundation Model Governance

Frontier AI systems pose a new challenge for global governance because of a number of features intrinsic to their development and deployment, and limitations on the resources available for governance today. The dominant approaches by governments and international organizations thus far, responding to governance challenges and ethical problems through principles and guidelines, have been limited in impact.

Development refers to the process of algorithm design to code machine learning models, followed by intensive model training on data using substantial compute resources to enable various AI capabilities. Deployment refers to bringing an AI model into a consumer or enterprise market.

Category	Challenge	Description
Development Challenges	1. Model Opacity	Foundation models trained with opaque techniques, 'black boxes' we do not understand.
	2. Unexpected Capabilities	Foundation models have new, unpredictable capabilities that go undetected during development.
	3. Limitations of Evaluation	Assessing a foundation model's capabilities cannot prove a model's safety.
Deployment Challenges	4. Misuse Potential	Foundation models pose societal scale misuse risks, such as enabling artificial pandemics, autonomous weapons, and widespread disinformation.
	5. Easy Proliferation	Once trained and released, foundation models can be easily proliferated, distributed cheaply across borders, copy-pasted to any computer.
	6. Dual-use Capability	Foundation models promise enormous political, military, and economic advantages for developers, which may compromise safety measures.
Governance and Resource	7. Rapid Development	Foundation models develop rapidly; models that were 'frontier' a year ago are both outdated and widely

if you can create a substitute or augmentation for human labour and sell it into the world, you are going to become more influential over time."

Limitations		accessible, requiring anticipatory governance.
	8. Autonomous Agents	Foundation model developers are aiming for super-intelligent, general, autonomous AI agents, which pose significant future global risks that must be considered now.
	9. Limited Safety Resources	Foundation AI safety receives a small fraction of scarce resources, limiting the number of researchers or computing power compared to the AI field as a whole.
	10. US Private Actor Dominance	Foundation model development is highly concentrated and dominated by a few AI firms, which is a challenge for regulation and leads to unequal access.

Development Challenges

1. **Model opacity:** Today the prevailing methods for training models are opaque, with current foundation models often described as 'black boxes.'[16] Even in the development phase, these models already conduct so many computations and become so complex that they lose interpretability – the ability for humans to readily understand the reasoning behind predictions and decisions made by the model.[17] Current attempts to develop insight into the models, sometimes termed "mechanistic interpretability" are not concrete actionable plans, but rather theoretical aspirations.[18]

2. **Unexpected capabilities:** Because of Model Opacity, new or 'emergent' capabilities are unpredictable and often go undetected during development and deployment.[19] For instance, Stable Diffusion was a foundation model designed to generate images. However, months after its release, external researchers demonstrated that the same model can be fine-tuned to produce music by converting sounds to images.[20]

[16] Stuart Russell said in a recent Senate Judiciary hearing that "They are black boxes…their internals are largely impossible to understand."https://www.judiciary.senate.gov/download/2023-07-26-testimony-russell

[17] Chris Olah of OpenAI noted that, as of 2021, the largest model that human scientists "really carefully understood" was around 50,000 parameters. Today's foundation models have trillions of parameters. https://80000hours.org/podcast/episodes/chris-olah-interpretability-research/

[18] Interpretability can be categorized as *mechanistic* (understanding the exact function of a system's subset), *concept-based* (identifying clusters of components that represent specific ideas or themes), and *feature-based* (pinpointing the importance of specific elements or attributes in the system's analysis). Despite this being a nascent field, it is still important to develop partial methods to reduce uncertainty in the model.

[19] Anthropic's CEO said that "You have to deploy [the model] to a million people before you discover some of the things that it can do."

[20]https://flowingdata.com/2022/12/16/stable-diffusion-spectrogram/#:~:text=Stable%20Diffusion%20is%20an%20AI,that%20is%20converted%20to%20audio.

Additionally, rather than following a linear trajectory, with capabilities improving steadily over time, frontier AI systems often progress through sudden leaps in functionality and capability as they scale or increase in size.[21]

3. **Limitations of evaluation:** 'Model evaluations' form the basis for AI licensing, auditing, and standards regimes. Model evaluations primarily consist of a group of evaluators prompting models to elicit certain capabilities and behaviours. While a good first step to notice and reduce risks, current techniques are primitive, and are unable to comprehensively assess foundation AI models. Even robust evaluations may fail to detect emergent capabilities arising post-deployment, or potential misalignments with human values.[22] Additionally, many model evaluations are conducted internally, which presents conflicts of interest.[23]

Deployment Challenges

4. **Misuse potential:** Foundation models have potentially societal-scale misuse risks. In the coming years, future AI systems may enable the widespread design of deadly pathogens, novel and possibly crippling cyberattacks, the propagation of persuasive disinformation, and the surveillance and suppression of dissidents, among other risks. The deployment of AI could disrupt nuclear stability or enable autonomous weapon systems.[24] This wide potential dual-use means that sector-specific regulations will be like a game of 'whack-a-mole,' leading to limited impact.

5. **Easy proliferation:** Once trained, foundation models can be copy-pasted and proliferate rapidly across borders. This makes global accountability and control both essential and difficult, especially if the models are open-sourced. In general, readily-trained models can be easily modified by anyone with access, including the removal of safeguards, and become a target for hijacking by adversaries.[25]

6. **Dual-use capability:** AI has massive dual-use potential. Foundation AI may be best thought of as a General Purpose Technology - a "single generic technology, recognizable as such… [that] comes to be widely used, to have many uses, and to have many spillover effects."[26] The potential benefits to be gained from controlling the most advanced foundation AI models are massive. In the United States, Goldman Sachs predicts AI could automate a quarter of current work, driving a global GDP increase of almost $7

[21]https://www.lesswrong.com/posts/5WECpYABCT62TJrhY/will-ai-undergo-discontinuous-progress#Defining_Discontinuous_Progress
[22] "Model evaluation for extreme risks" https://arxiv.org/abs/2305.15324
[23] https://arxiv.org/abs/2001.00973
[24] https://arxiv.org/pdf/2307.03718.pdf
[25] https://arxiv.org/pdf/2307.03718.pdf
[26] Economic transformations: general purpose technologies and long-term economic growth. OUP Oxford, 2005.

trillion.[27] This could encourage AI companies to accept a "safety race to the bottom," whereby they are disincentivised to follow any industry best practices that slow down their rate of development. They may be further willing to absorb short-term losses to establish monopolies, which would then lead to increased wealth gaps and social unrest without systems for redistribution.

Governance and Resource Limitations

7. **Rapid development:** The speed of AI advancements continues to exceed expectations. According to current trends, every year, there is an approximately 6–20x increase in the capabilities of models, arising from the multiplicative effect of improvements in hardware quality, quantity, and algorithmic progress.[28] Models that were 'frontier' a year ago are now both outdated and widely accessible, and this trend is likely to continue.[29] Though still nascent, there is growing evidence that a recursive process of AI research can help the efficiency of AI research, reducing data centre energy costs, improving the efficiency of programmers, enhancing semiconductor design, and producing language models that fine-tune on their own output.[30] Leading scientists predict that human, and super-human level AI systems, could be developed in the next two decades, and potentially in the next few years.[31] This means that AI, like some other emerging technologies, suffers from the "Collingridge dilemma," whereby regulators who wish to prevent harm from new technologies must create norms and regulations before the potential impact of technology or their regulations are known.[32]

---

[27] https://www.goldmansachs.com/intelligence/pages/generative-ai-could-raise-global-gdp-by-7-percent.html

[28] Further, the progress is predictable and driven by some simple underlying factors that are not likely to slow down anytime soon. Specifically, the power or intelligence of an AI system can be measured roughly by multiplying together three things: (1) the quantity of chips used to train it, (2) the speed of those chips, and (3) the effectiveness of the algorithms used to train it. The quantity of chips used to train a model is increasing by 2–5x per year. Speed of chips is increasing by 2x every 1–2 years. And algorithmic efficiency is increasing by roughly 2x per year. These compound with each other to produce a staggering rate of progress. Dario Amodei, CEO of Anthropic, written testimony to the Senate Judiciary. https://www.judiciary.senate.gov/imo/media/doc/2023-07-26_-_testimony_-_amodei.pdf

[29] In 2020, experts predicted AI wouldn't pass SAT exams until 2057. By 2023, they consistently get top scores. In the same few years, AI went from being barely being able to read and write to creating award-winning photographs and art, convincingly cloning voices in seconds, and deceiving people into thinking they're human.

[30] https://www.deepmind.com/blog/deepmind-ai-reduces-google-data-centre-cooling-bill-by-40, https://arxiv.org/abs/2110.11346., https://arxiv.org/abs/2207.14502.,

[31] "I and other leading AI scientists now believe human-level AI could be developed within the next two decades, and possibly within the next few years… The shorter timeframe, say within 5 years, is particularly worrisome because scientists, regulators and international organizations will most likely require a significantly longer timeframe to effectively mitigate the potentially significant threats to democracy, national security, and our collective future." Yoshua Bengio, "[...]Once we can develop AI systems based on principles akin to those underlying human intelligence, these systems will likely surpass human intelligence in most cognitive tasks, i.e., we will have superhuman AIs."

[32] https://www.sciencedirect.com/science/article/pii/S0048733317301622

8. **Autonomous agents:** There might be significant risks from large, autonomous systems that we cannot yet know. As systems like AutoGPT illustrate, it is straightforward to turn existing Large Language Models (LLMs) into autonomous agents, those capable of "autonomous, purposeful action in the real world."[33] Today, programmes like OpenAI's plug-ins allow LLMs to control third-party applications, allowing the model to order products, send emails, and browse the web. We do not know how to control autonomous, self-improving systems that leading companies currently have ambitions to build.[34] It is difficult for policymakers to weigh the enormous benefits of AI today against these risks.

9. **Limited safety resources:** The technical AI field itself suffers from a shortage of safety researchers, with estimates ranging from 100–500 researchers who are purely focused on safety across academia, private labs, and non-profits. Furthermore, the technical problem is difficult. Despite many years of work on how to make safe systems, there has been little theoretical or empirical progress. We can neither predict the range of possible capabilities large systems can develop, nor the limits of their existing ones. We should invest at least as much in research to protect the public from powerful AI systems as we are globally investing in increasing their capabilities.

10. **US private actor dominance:** As this development process requires extraordinary amounts of talent, hardware, and data, state-of-the-art foundation models can currently be developed by just a few laboratories powered by three big cloud providers located in the United States. Countries directly overseeing frontier AI systems may be hesitant to relinquish authority or comply with global norms or regulations.

[33] https://iiif.library.cmu.edu/file/Newell_box00089_fld06093_doc0001/Newell_box00089_fld06093_doc0001.pdf

[34] Scientists cannot predict the result of an arbitrary prompt into a chatbot, but we can put bounds on its output, such as limiting the response to below maximum allowed output. But once this system starts looping, and starts to leverage external systems, those break, and we can't reliably predict or control what happens. Some AI companies are moving in this direction, but regulators will struggle to respond. Models are being tested in virtual environments like Minecraft, and OpenAI is seeking $100 billion to explicitly build recursively improving AI.

AI as 'Tragedy of the Commons': If Developers are Worried About Foundation Models, Why are They Building Them?

State-of-the-art foundation models are currently being developed by a few major AI labs – OpenAI, Google Deepmind, and Anthropic – followed by others such as Meta, x.ai, and Inflection. Most of these major AI labs, and their leaders, have stated on record that they are worried about the harms and risks, but still continue to push ahead. This has led some to suspect ulterior motives: companies might be highlighting extreme risks to amplify their own importance in regulatory discussions; to market the power of their products; or to deflect responsibility for more short-term harms such as misinformation.

However, hundreds of prominent independent academics, such as Turing Award winner Yoshua Bengio, have also signed a recent statement voicing concerns about AI risk. Prof. Stephen Hawking already warned in 2014: "The development of full artificial intelligence could spell the end of the human race." Several of the founders of frontier AI labs warned the world about the existential risk posed by AI before the founding of their labs.[35] Sam Altman, for instance, takes no equity in OpenAI, possibly to avoid having an incentive to push ahead recklessly. If they are so worried, why don't they stop?

This situation can be explained aptly by the "tragedy of the commons," rather than simple competition. In the frontier AI context, the 'commons' are society's capacity to benefit from AI without tipping into disaster. Societies and their institutions can only adapt to technological breakthroughs at limited speed. When technological advancements outpace society's capacity to adapt, severe harms can result across scales – from the individual to the existential.

The tragedy arises because in the absence of regulation, unilateral restraint is futile. If one AI developer slows research, competitors will continue to charge ahead, creating incentives to develop further, despite the risks. AI company safety teams play an important role but suffer from conflicts of interest. Thus, effective **regulatory action is needed for all developers to accept a slow-down**. Self-regulation is insufficient when rewards are immediate but harms remote. Only binding oversight can coordinate responsible progress in a dispersed ecosystem where individuals operate locally while impacts are global.

[35] Shane Legg, in 2011, wrote that AI risk was his "Number 1 risk for this century, with an engineered biological pathogen coming a close second." In 2015, Sam Altman, founder of OpenAI, wrote: "Development of superhuman machine intelligence (SMI) is probably the greatest threat to the continued existence of humanity."

Part II:
Assessment of Proposed International Institutions in AI Governance

Why It Seems Premature to Launch an International Organization for AI

Overview Part II:

I. Abstract:

Recently, there have been numerous proposals for new international institutions to govern AI risks, especially AI foundation models. The suggested approach in this section draws inspiration from existing bodies like the International Atomic Energy Agency (IAEA), the European Organization for Nuclear Research (CERN), the International Civil Aviation Organization (ICAO), and the Intergovernmental Panel on Climate Change (IPCC).

Th section of the report analyses the potential strengths and limitations of each institutional model for governing foundation models.

Key findings:

- The IAEA model offers useful verification mechanisms but would face challenges given the lack of consensus on standards for AI systems and the inability to match the rapid pace of AI progress.
- A CERN model could enable beneficial collaboration but seems mismatched as a governance structure and could hinder decentralized innovation in AI safety research.
- The ICAO allows for harmonized standards but achieving global treaties will be challenging in the current geopolitical climate and it likely cannot respond rapidly enough to exponential AI advances.
- The IPCC model can build consensus and offer credible guidance but lacks enforcement capabilities and may divert limited AI expertise away from other efforts.

Overall, while existing institutional models have shortcomings around enforceability, responsiveness, and scope, they offer historical precedents to inform future AI governance. This examination shows that any international AI regime will leverage current institutions and require a multifaceted strategy. Global oversight will not arise in a vacuum but rather build on lessons and frameworks developed over time. A nuanced, multi-pronged approach harnessing the strengths of various models will prove most effective in managing a complex technology spanning borders and sectors.

II. Overview of Proposed Institutional Models

Currently, many ideas in international AI governance are centred around establishing an institutional global governance response to manage the accelerating development of frontier AI systems.

Prominent AI researchers and leaders, including OpenAI CEO Sam Altman, have publicly advocated for an AI governance model similar to the International Atomic Energy Agency (IAEA). These calls were most recently echoed by UN Secretary-General Guterres in comments to the press in June 2023. DeepMind researcher Lewis Ho elaborates, proposing an 'Advanced AI Governance Agency,' modelled after the IAEA, for rule-making and enforcement on the international level. This agency would advocate for the adoption of standards and norms, assist in their execution, and monitor adherence, thereby guiding the responsible use and development of advanced AI.

A multilateral effort to develop and build advanced AI was first proposed by Prof. Gary Marcus in 2017. More recently, Ian Hogarth, the current Chair of the UK foundation model taskforce, suggested a thought experiment for safe AI development known as "The Island," whereby a highly secure facility would build advanced AI, and all other development would become illegal. Ho et. al. further considered the establishment of a project akin to CERN for AI safety. The creation of a 'Frontier AI Collaborative,' designed to parallel the GAVI vaccine alliance, would focus on the advancement, distribution, and access to frontier AI technologies, paving the way for widespread AI adoption. This initiative would carry out research in AI safety, thus promoting safer applications of AI technologies.

Other proposals – such as the "International AI Organization (IAIO)" – function similarly to the International Civil Aviation Organization (ICAO): starting with soft law instruments and eventually formalizing its regulatory role. This is similar to an idea by Wallach and Marchant for the international oversight of AI governance, performing coordination, monitoring, analysis and convening functions. The need for harmonized standards has been widely recognized. The UN's Global Digital Compact policy brief also calls for "sector-based guidelines to ensure that technology developers and other users have applicable, relatable guidance for the design, implementation, and audit of AI-derived tools in specific settings." Similarly, the G7 declared its support in May of this year for "the development and adoption of international technical standards in standards development organisations through multi-stakeholder processes" as part of its "Hiroshima AI Process."

Finally, in Secretary-General Guterres's policy brief on a Global Digital Compact, he calls for a "global, multidisciplinary conversation in order to examine, assess, and align the application of AI and other emerging technologies," and a "need to bring stakeholders together in a meaningful effort to consider the implications of emerging technologies and ensure that they align with universal human rights and values." The IPCC model is hypothesized to achieve this through its ability to pool knowledge, build consensus from a diverse group of experts, and communicate

evidence-based policy effectively to policymakers and the general public. A *Nature* article offered a detailed blueprint for what such an IPCC, expert-led observatory could entail. In their submission to the UN Secretary-General's High-level Panel on Digital Cooperation, research groups at Oxford and Cambridge also endorsed this model for the UN to "provide a legitimate, authoritative voice on the state and trends of AI technologies."

While there is a wealth of novel ideas and principles for global AI governance, they all share common shortfalls. None evaluate what needs to be done in the immediate short-term (up to 2024) to address the unfettered growth of foundation models. There is also little consideration for the Global South, and how to ensure that the benefits from AI are evenly distributed. Finally, there are gaps in the feasibility of implementing any such regime, including an examination of the strengths and weaknesses of the UN in light of an exponentially developing technology. These shortcomings are discussed in more detail below but are listed here for ease of reference.

It is clear that there is no singular perfect solution. A multipronged approach will be necessary to regulate AI at various levels to both tackle urgent, near-term issues, as well as start establishing institutions for future governance.

Model	Strengths as a model	Weaknesses as a model
IAEA	- Proven success with nuclear technology - Established verification mechanisms - Some analogies, such as AI hardware and uranium	- Challenges verifying opaque AI systems - Limited safety expertise of the UN - Unable to match pace of AI progress - Focused on States, not companies
CERN	- Enables large-scale collaboration and benefit-sharing - Could aggregate safety research	- Does not directly address governance, instead a joint scientific endeavour - Proliferation issues remain - Centralization of AI safety may limit speed - Highly ambitious, many short-term implementation challenges - Difficulty in persuading all actors to join

ICAO	- Respects national sovereignty and standard-setting capabilities - Can harmonize standards - Legal authority via treaties	- Limited practicality of achieving a treaty in current geopolitical landscape - Unable to match pace of AI progress - US private sector not included
IPCC	- Expert consensus-building - Advisory capacity - Multilateral credibility	- Lacks enforcement authority - Lag between research and policy - Advanced AI may be on a shorter timeline than climate change

IAEA – Standards and Compliance

The International Atomic Energy Agency (IAEA) conducts regular inspections and monitoring to verify member states' compliance with their commitments under international treaties and agreements. The IAEA ensures that nuclear materials are not diverted from peaceful uses to military purposes. Proponents look to the IAEA as a model of how 1) the world can cooperate and agree on common standards, and 2) how verification of compliance to international standards may be effectively enacted.

The model has an understandable appeal due to its:

- **Success:** The IAEA has contributed to preventing nuclear conflict for over 50 years. It has limited proliferation to just nine countries while facilitating the development of safe nuclear power to 33.

- **Ostensible technological similarities:** Both nuclear power and AI are defined by their 'dual-use' capacity – the potential for mass harm and mass benefit. Further, race dynamics are often invoked to justify their unchecked form of development.

- **Practical feasibility:** There is a clear analogy that could track and monitor AI development via hardware. Large AI models currently require highly specialized chips, whose supply chain is concentrated in a few countries that could be tracked, similar to uranium production.

However, the IAEA model has serious limitations. Unlike nuclear weapons, the scope of applications for foundation models, a general-purpose technology that can be deployed across society, is much greater. Advancing an IAEA model for AI seems premature due to challenges with:

- **Verification:** IAEA-style verification requires clear standards which can be used to audit and verify compliance, but there is no consensus yet on what these standards should be for AI systems. One option could include inspecting hardware, requiring countries to allow direct physical access to data centres, which could be difficult to obtain. Another option is to use model evaluations after development, which involve running tests to assess the competencies of the model. But, this faces limitations as a post-hoc intervention. Agreeing on any best practices for verification will not be trivial from either a technical or policy perspective.

- **Limited safety expertise:** If technical standards for verification are agreed upon, there will then be a need for highly competent experts to conduct evaluations in a comprehensive and timely manner. Currently, model evaluations are being conducted by just a few dozen leading experts. Training and attracting talent to serve as auditors will be costly and time-intensive. Additionally, as AI systems proliferate, the number of verifiers needed to mitigate AI risks may grow exponentially.

- **Speed:** An IAEA is unlikely to be able to respond to the rapid development of Foundation Models. The IAEA model may be good at tracking the use or development of known technology, but not one that grows in exponential ways. For example, the IAEA coordinates research projects to keep pace with state-of-the-art technologies. A coordinated research project, at best, produces results after 1–2 years and at current development speeds, the state-of-the-art in AI would already have advanced significantly.

- **Subnational actors:** The IAEA functions on a State level, and its model is not directly suited towards subnational actors. Compared to State weapons programmes, sub-national actors like technology companies remain dominant in the development and deployment of foundation models. Furthermore, given US Private Actor Dominance, an IAEA model will require the buy-in of the United States, who would likely prefer their own national regulators and auditors to be supported in a first step, rather than immediately submitting to international ones.

CERN – A Collective Scientific Endeavour

CERN, or the European Organization for Nuclear Research, is unique in that it collectively provides hardware for particle collision. The institution created a collective scientific endeavour to enable complicated and expensive foundational research. CERN's creation was not about risk mitigation, but rather joint scientific advancement made possible through collaboration.

There are two distinct institutions referenced when the concept of a 'CERN-for-AI' is invoked:

(1) Joint AI safety research: resources of signatory States are centralized for the development of techniques to make frontier AI (developed outside of the institution) safer. This includes research on how to determine the capabilities of these systems in advance, make these systems more interpretable, and detect and prevent dangerous capabilities such as deception.

Type (1) increases the scale, resourcing and coordination of AI safety research for the technical mitigation of risk. The institution would need significant computing resources and technical expertise, and access to the latest foundation models. Leading labs have already demonstrated their willingness to cooperate with regulators, such as by sharing access with the UK Government and making voluntary commitments to conduct internal security assessments before release, and sharing information with the US Government.

(2) Joint advanced AI research: resources of signatory States are centralized to research frontier AI, with some proposing that all other AI development become illegal. This lab would be exclusive, the world's only facility permitted to conduct advanced AI foundation models research, focused on developing safe architectures, being highly secure, and distributing the benefits of its research to participating members.

If implemented effectively, type (2) goes even further in mitigating AI risk: the requirement for exclusivity, enforced via a moratorium on models exceeding a certain threshold, means that any future advanced AI system is developed and deployed as a shared human endeavour, in the facility and nowhere else.

However, either 'CERN-for-AI' model would likely fall short for a combination of practical and political reasons:

- **Model mismatch:** Fundamentally, CERN is not a governance structure. CERN was created because none of its founder countries had the capabilities to pursue advanced nuclear physics research. Foundation model development is currently dominated by large companies, not bottlenecked by national capabilities. CERN also did not prohibit the construction of other particle accelerators. Its structure is, therefore, more about common scientific endeavour, rather than limitation of risk.

- **Proliferation:** In the long-term, the easy proliferation of frontier AI models makes control difficult. If no accompanying measures are implemented, by the time that a shared, exclusive compute cluster is created and enforced, by all expectations, frontier AI may be accessible on consumer-grade hardware. If the CERN was created to conduct experiments that no single State could do on their own, AI foundation models will soon be possible for a much wider range of actors.

- **Limits of centralization:** Centralization of AI safety research may limit innovation. A diversity of approaches and research groups would likely yield faster progress. Additionally, while hardware resources could be consolidated, it would be unnecessary to physically relocate researchers to a single location. A more adaptable approach could involve a small (<10) network of coordinated labs with independent scientific direction.

- **Practical problems:** Beyond the limited technical expertise, in AI safety and within the United Nations, both models have challenging implementation problems. Firstly, 'CERN-for-AI' may pull away safety researchers and struggle to get model access from leading labs. Secondly, it is contingent on the buy-in of major AI developers and the jurisdictions that host them, primarily the United States.

- **Incentive alignment:** Frontier AI is developed separately by private actors with profit motives. There is little incentive for them to develop frontier AI systems together. This scenario is more like an arms race than a collective endeavour, so building an institution that assumes common scientific purpose seems unrealistic.

ICAO – Harmonizing and Internationalizing Standards

The International Civil Aviation Organization (ICAO) is an example of a specialized UN agency that leverages expertise to set internationally-recognized aviation standards, promote their implementation, and oversee countries' compliance. Out of the four models explored, the ICAO is the only one that is a UN-operated agency. It draws its authority from the Chicago Convention, an international treaty adopted by virtually all countries to regulate aviation and air space.

Unlike the IAEA, the ICAO does not possess direct enforcement powers. It relies on Member States to implement and enforce the standards and regulations within their respective jurisdictions, and is therefore less intrusive on State sovereignty. It does, however, depend on incentive structures (both carrots and sticks) for States to comply with the standards.

The ICAO offers many advantages:

- **National sovereignty:** This is the key advantage of the ICAO model – its respect for the sovereignty and expertise of national agencies, which may be attractive to powerful nations resistant to global oversight.

- **Interoperability:** It reduces cross-border frictions and the burden on domestic regulators, especially in smaller countries with less technical expertise, to identify necessary safety protocols.

- **International regulation:** In theory, an ICAO model could lead to restrictions on countries that are not certified, in the same way that Member States restrict flights from jurisdictions without ICAO certification from entering their airspace.

- **Legal enforcement:** In an era of high-risk, legal authority matters, the ICAO's near-universal ratification and binding enforcement, while also respecting State sovereignty, present key advantages for effective global governance.

The ICAO model, however, presents difficulties in terms of:

- **Feasibility:** It seems extraordinarily difficult to achieve a global treaty on AI governance in the current geopolitical landscape. It would also be premature to announce a large-scale standards body with sufficient buy-in from key stakeholders,[36] given the lack of consensus on best practices for monitoring AI models (see limitations of evaluations).

[36] The US and China jointly defining standards would set a powerful precedent for multilateral efforts.

- **Speed:** Recent efforts to achieve international treaties, for example on migration and ocean governance, have either failed or required a gruelling, multi-year process. Foundation model development requires rapid and dynamic, not static, governance. Models that were 'frontier' a year ago are now both outdated and widely accessible, and this trend is likely to continue. The ICAO's lengthy consensus-based process takes on average two years for a new standards proposal to be formally adopted. This is not sufficient for the rapid pace at which AI has been developing.

- **Model mismatch:** The ICAO primarily focuses on governing the civil aviation operations of planes *after* they have been developed, and the physical machinery with well-established engineering principles with little innovation. In the case of frontier AI, a lot of the risk has already emerged during research and development. For ICAO, regulatory standards and guidelines apply mainly to the operation and certification stages, with a focus on safety, efficiency, and environmental impact.

- **Static:** Unlike aviation, where physical infrastructure and operations do not change fundamentally year-on-year, AI systems allow for easy proliferation and are rapidly evolving. It is difficult to agree on sufficiently concrete technical guardrails in a treaty when the technical risks and responses are constantly evolving.

- **Subnational actors:** Beyond States, an AI treaty would also need to include industry, for which there is little precedence, though one recent example is the plastics treaty negotiations, which counted 190 industry representatives at the Paris talks. A model based solely on State agreements is unlikely to be effective in AI.

IPCC – Expert Knowledge-gathering Panel

Like climate change, AI has unpredictable consequences that cross generations and borders, leading numerous researchers to propose a global AI observatory similar to the Intergovernmental Panel on Climate Change (IPCC).

Of the four models examined, the IPCC model seems the most promising as a first step in global foundation AI governance. In our recommendations, we propose a similar, scaled-down version for a new international institution.

The IPCC is an intergovernmental body formed through cooperative resolutions of WMO and United Nations Environment Programme (UNEP) Member States. It derives its authority from a rigorous scientific process, rather than any legal treaty. The IPCC principally serves as an advisory body of scientists tasked with collecting and collating scientific consensus on issues related to climate change. It then offers policy relevant recommendations which carry weight due to its intergovernmental approach.

An IPCC-like governance model for AI could be advantageous in several ways:

- **Consensus-building:** It could act as a **steward of knowledge** on frontier AI systems, build a registry of existing models, establish a shared body of data and analysis, forecast trends, and orchestrate global forums to debate advancements in the field.

- **Education:** It could serve an educational function, broadcasting technical knowledge in simple terms and explaining counterintuitive concepts, such as why open source is likely not the best approach for frontier AI development. It would also incentivize more research in academia and elsewhere, thus increasing the number of experts.

- **Scientific rigour:** The panel's reports could derive their credibility principally from an extensive, transparent, and iterative peer review process. Although the size and complexity of its review process could undermine its ability to respond rapidly to emerging advances in the field.

- **G77 credibility:** It could gain wider legitimacy in non-Western countries – particularly China – and allow for wide representation among developing countries.

- **Inexpensive:** In 2022, the IPCC's proposed annual budget was approximately $6 million.[37] An 'IPCC for AI' could operate with a similar annual budget.

However, the IPCC model for AI may struggle to overcome:

- **Lack of teeth/enforceability:** The IPCC lacks any 'hard' authority to monitor, verify, and enforce compliance for advanced AI standards. Its recommendations could simply be dismissed, and thus it may not be powerful or comprehensive enough to mitigate wide-scale societal risks from advanced AI.

- **Divergent risk frameworks:** There is a lack of consensus within the broader AI research community regarding the level of risks potentially posed by AI. Many scientists hold divergent perspectives on the potential for existential catastrophe, believing that these dangers, while reasonable, seem too hypothetical to warrant investment of limited resources. Absent shared understanding of the core hazards, it becomes difficult to motivate action or alignment on safety goals. Bridging divides over risk perception and priorities would be a necessary component to the agenda of this body.

- **Lag:** Scientific consensus can reach maturation far before meaningful global policy action. Despite widespread scientific consensus, climate action, and the IPCC, took decades to gain credibility. An 'IPCC-for-AI' may suffer similarly. AI may require an even shorter timeline than climate change, given that a growing number of experts working directly on leading models believe highly advanced AI systems are possible within the

[37] Chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://apps.ipcc.ch/eventmanager/documents/71/151020210812-Doc.%202%20-%20IPCC%20Programme%20and%20Budget.pdf

next few years.

- **State ownership**: While State authority lends legitimacy to the IPCC model, it can also allow governments to influence the outcomes, for example by watering down recommendations. Additionally, AI oversight is largely outside of State control. Private sector developers may dismiss State-centric guidance. Governing foundation AI models requires more flexibility than this model currently permits.

- **Opportunity cost:** There are very few AI experts worldwide focused on safety. As most are already heavily engaged in national processes, the UN would be competing for technical talent and diverting attention from domestic regulation efforts. It may better serve by facilitating an informed international discourse with a much-needed focus on foundation models.

- **Closed doors:** Scientific research by an 'IPCC-for-AI' would be significantly hindered by the lack of transparency from foundation model providers, who rarely disclose meaningful insights regarding the data, compute, and deployment of their models, nor the key attributes of the models themselves.

Part III:
A Framework for Avoiding the AI Tragedy of the Commons

Guidance for The Advisory Body

Overview Part III:

I. Abstract

In this section, we assess the UN's strengths and limitations regarding international AI governance and emphasize that any UN international institution charged with AI governance should focus on norm-building rather than a hard regulatory mechanism. In its deliberations on the exact form this institution should take, the Multistakeholder Advisory Body on AI should engage the private sector directly in the UN's multilateral AI efforts and ensure harmonization with national AI regulatory processes, particularly in the US and China.

We present our recommendations through Nobel laureate Elinor Ostrom's eight-step framework for avoiding the "tragedy of the commons" – the commons being society's capacity to benefit from AI without tipping into disaster. Through the framework's eight steps, we derive the immediate next steps the Advisory Body can take to leverage the UN's capabilities for multilateral AI governance.

Key Findings:

- Any UN AI institution should focus on moral authority, not technical proficiency. The UN can effectively promote norms and inclusion, given its global platform. However, its State-centric composition, lack of technical expertise, bureaucratic lags, and limited enforcement authority over private developers curb its regulatory oversight capabilities in AI.
- The rapid development of AI poses risks of a "tragedy of the commons" if left ungoverned, threatening humanity's collective capacity to adapt to rapid technological change. The UN is uniquely placed as a global institution to respond to this dilemma.
- The Advisory Body should propose an international regime which amplifies the voices of less powerful actors, advocates for equitable distribution of benefits, and builds consensus around universal norms within AI.

II. Assessing the UN's role in AI Governance

The challenges common across all of these models point to the difficulties in managing the global risks posed by the rapid evolution of AI.

There are a growing number of calls from all nations, including those within the Global South, to regulate the development of foundation AI. The Chinese Ambassador to the UN, Zhang Jun, set out China's vision at the UN Security Council Briefing on Artificial Intelligence, highlighting that we must ensure "risks beyond human control do not occur," and that "mankind has the ability to press the stop button at critical moments." The Deputy Minister for Foreign Affairs for Mozambique, Manuel Gonçalves, said at that same meeting: "In the event that credible evidence emerges indicates that AI poses an existential risk, it's crucial to negotiate an intergovernmental treaty to govern and monitor its use."

Moreover, several leaders within the tech industry have called for a UN role in regulation. For example, Jack Clark of Anthropic recently addressed the Security Council and called for international regulation of big tech firms. Sam Altman of OpenAI has called for an IAEA-like institution that would need to be administered globally.

While there is a clear demand for international oversight, it is less clear that the UN should be the designated authority to enact it. Indeed, today foundation AI development takes place entirely outside of the UN system, and there is no immediate mandate within the UN Charter to address AI per se. However, as the Secretary-General has emphasized, the UN Charter's call to "protect succeeding generations" gives the UN a mandate to address existential risks, and the UN may be an "ideal place" to develop global standards and risk mitigation as a globally representative entity.

The recommendations below are targeted towards the agenda of the nascent Multi Stakeholder Advisory Body on AI and could apply across the entire UN system, including the United Nations General Assembly (UNGA), programmes and funds, the United Nations Economic and Social Council (ECOSOC), specialized agencies, the United Nations Secretariat and offices, and the United Nations Security Council (UNSC).

Strengths

There are compelling advantages for a UN role in supporting the development of AI governance standards and regulation:

- **Establishing global norms in AI:** The UN has a long and illustrious history of crystallizing global norms, such as human rights. For example, the Outer Space Treaty (1967), facilitated by the UN, established non-armament norms in space, demonstrating the organization's ability to successfully guide international behaviour. Its enforcement, though primarily reliant on diplomatic and political pressures, has held strong over

decades, underlining the UN's role in maintaining global standards. In the case of AI, this could concern common practices around the responsible development and deployment of AI systems. The UN could offer a valid framework for universal AI norms, giving global AI risks a platform for legitimacy.

- **Safeguarding humanity:** The UN has a proven record in equalizing global inequalities, particularly seen in the adoption and delivery application of the Sustainable Development Goals (SDGs). This could be instrumental in managing AI resources and reducing the risk of unequal application. For foundation models deemed safe, the UN could facilitate knowledge sharing across different sectors, and (as the United Nations Development Programme does) encourage technologically advanced countries to provide AI safety training for less-developed nations.

    - **Disproportionate risks:** Global South countries may face unique vulnerabilities in the face of powerful frontier AI models. For example, the scale and complexity of potential cyber threats may overwhelm their typically limited cybersecurity infrastructure. Economic coercion, information control, and information manipulation may be used to further sway public opinion or disrupt political processes and stability.

    - **Consensus building:** Global South countries are generally underrepresented in global governance. The UN remains a unique forum where the voices of all nations, regardless of their size or power, can be heard. This is essential in ensuring that the impact and benefits of AI are universally shared. Furthermore, democratic participation is intrinsically valuable, and Global South countries currently participate less in the AI debate, so amplifying their voices through the UN makes the debate more democratic.

    - **Guidance for under-resourced nations:** Given that there is already evidence that Global South Member States are concerned about the risks from AI, the UN could offer assistance to help smaller countries navigate effective global AI regulation. The UN could offer crucial direction to smaller, technologically underdeveloped countries, helping to foster more balanced AI advancement across the globe. These countries are fully exposed to these risks but do not have direct benefits in the same way as, for example, the United States does. Hence, if technologically under-developed countries were sufficiently informed, they would likely act more prudently on this issue.

Limitations

There are a number of challenges that the present UN system needs to overcome:

- **Limited resources and technical expertise:** The UN faces a severe shortage of resources and technical AI expertise needed to assess frontier AI risks. Between 1980 and 2015, only 2 per cent of UN agendas pertained to science and technology while 83 per cent pertained to human rights, development, security, and governance measures.[38] This shortage also extends to most governments and the field of AI safety itself. In 2022, it was estimated that there were roughly 100 researchers dedicated to AI safety, while there were over 100,000 researchers working on furthering AI capabilities.[39] Global investment in AI systems has been forecasted to reach $154 billion by the end of 2023.[40] For public sector governance actors, including the UN, this imbalance of technical expertise and financial resources is not trivial to overcome. It also indicates the opportunity cost of diffusing researchers to various projects.

- **Enforcement challenges:** Enforcement of safety compliance within frontier AI systems is a technical issue with no current consensus on best practices. The success of modern AI techniques relies on computation on a scale unimaginable even a few years ago. Some proposals, collectively known as 'compute governance,' seek to capitalize on the fact that the large training runs required for frontier models utilize vast quantities of energy, providing clear footprints for regulators. Other proposals seek to monitor and control hardware, as frontier AI models require specific 'AI accelerators,' which are advanced computer chips specialized for AI applications. Currently, however, these chips are globally available and untracked. Overcoming these technical bottlenecks necessarily precedes policy action. Without concrete levers to ensure AI safety compliance, the UN risks launching an institution with no clear mandate or course of action, which could distract from more substantial regulatory processes at the national level. Any premature overreach in a regulatory capacity could also lead to the UN's dismissal by critical actors.

- **State-centric approach and non-State actors:** With AI development primarily driven by non-State actors such as technology companies and research institutions, the UN's State-centric composition structurally excludes those directly in control of AI systems.[41] This could change if national governments impose direct regulatory authority over the companies within their jurisdictions, as in the case of the European Union,

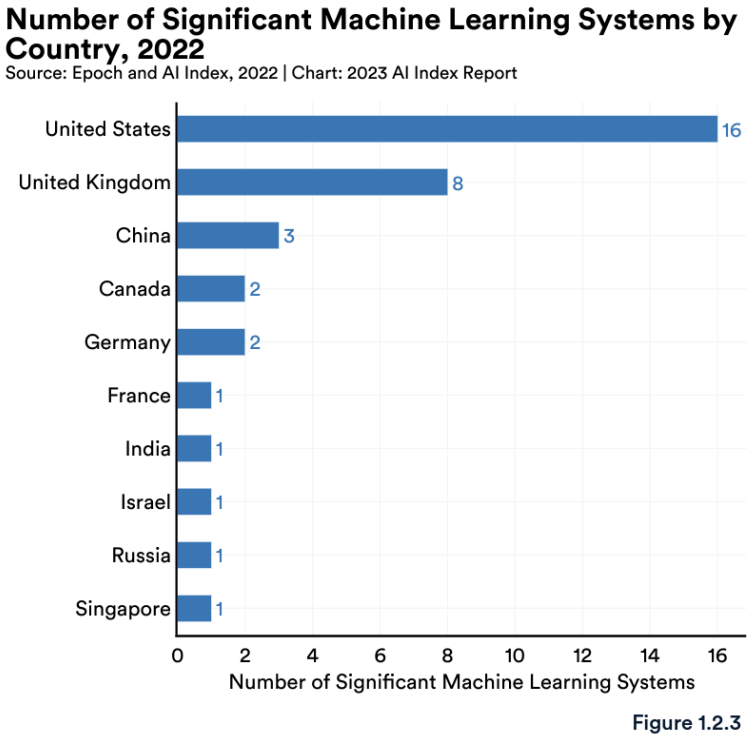
[38] https://link.springer.com/article/10.1007/s11558-017-9288-x
[39] https://www.lesswrong.com/posts/mC3oeq62DWeqxiNBx/estimating-the-current-and-future-number-of-ai-safety#:~:text=A%20recent%20post%20(2022)%20on,time%20on%20technical%20AI%20safety.
[40] https://www.idc.com/getdoc.jsp?containerId=prUS50454123
[41] It should be noted that state-centrism, however, could be advantageous in engaging China and other powerful states which prioritise state sovereignty.

though it seems unlikely in the short term.[42]

- **Geographic concentration:** Of these non-state actors, the majority are located in the US and UK (see table below). It appears currently difficult for the UN, which represents 193 Member States, to justify its jurisdiction when a smaller, more nimble unilateral, bilateral, or 'minilateral' approach could effectively cover the vast majority of advanced AI operations. On the other hand, it could be argued that the UN is very well positioned to put pressure on leading labs and the countries they're based in, since the majority of UN countries don't hold a stake in the success of these specific labs. Encouragingly, many current UN agencies and programmes allow for flexibility in their approach to engage with the private sector. Governing AI effectively will also mean involving non-State actors who will be key to implementing effective AI governance.[43]

- **Regulatory capture:** While not a problem specific to the UN, preventing regulatory capture[44] within these negotiations would additionally be the responsibility of the Multistakeholder Advisory Body on AI or other UN entity hosting them. This is important when considering recommendations for regulation that may not serve business interests. For example, China's call for a temporary stop if there is sufficient evidence that these systems are dangerous, may not be taken seriously by a panel of experts from AI companies.

[42] The same issue applies to climate change, where the main CO_2 emitters are private sector companies. Still, the UN has deployed a large regime complex with fora, expert bodies, and agreements to tackle the issue and has arguably achieved progress in curbing climate change.
[43] The Montreal Protocol negotiations in the 1980s provide one case study for the inclusion of private actors alongside national delegations. DuPont and other chemical companies provided technical expertise on the feasibility and costs of shifting away from CFCs and towards alternative chemicals. However, they also negotiated and lobbied for longer phase-out timelines that would be less disruptive to their businesses.
[44] https://www.sciencedirect.com/science/article/pii/S0016328721001695



Epoch and AI Index, 2022 | Chart: 2023 Stanford HAI Index Report[45]

- **Slow processes:** A survey of AI researchers in 2022 estimated there is a 50 per cent chance of human-level AI by 2057.[46] Since then, this timeline has been shortened to closer to 2030.[47] AI developments are built on path dependencies, and intervening early will help set the right trajectories to mitigate risks later on. The UN's processes, known for their broad principles and gradualist approach, may not be sufficient to keep up with AI's rapid pace. The Montreal Protocol, while a positive example of private sector engagement, took over a decade to implement a complete phaseout of ozone-depleting substances. This lag could lead to issues of adaptability in responding to the frontier dynamics of AI development. The multifaceted and often diverging priorities of the UN and its Member States could result in attention drift, or in the worst case, total paralysis in AI action.

- **Multilateral challenges:** The multilateral system is already under considerable strain. In recent years, the UN system has been called upon to address unprecedented and interlocking challenges including global pandemics, climate change, and growing geopolitical divisions, often referred to as a 'polycrisis.'[48] Preventing future shocks

[45] https://aiindex.stanford.edu/report/
[46] https://ourworldindata.org/ai-timelines
[47] https://www.metaculus.com/questions/4815/date-of-first-agi-according-to-forecasters/
[48] https://www.weforum.org/agenda/2023/01/polycrisis-global-risks-report-cost-of-living/

requires multilateral institutions to reorient from crisis response to foresight, anticipatory action, and agility, which they may not be designed to do.[49] Simultaneously, many countries are turning away from a 'rules-based order' and acting unilaterally or engaging in 'forum shopping,' rather than meaningful compromise.[50] Geopolitical tensions and polarization further impede the UN's ability to achieve rapid and effective response to AI's immediate dangers.

III. Eight Steps to Avoid the Tragedy of the Commons

Step One: Defining the Commons

Unless a specified good and community of benefit are defined and institutionalised, AI development will continue in a scenario of privatized gains and socialized losses.

In the case of AI development, the common good is society's capacity to benefit from AI without tipping into disaster. An example of a similar commons is the capacity of the Earth's atmosphere to absorb greenhouse gas emissions without tipping into climate disaster. It is a collective action problem, in which individual actors prioritize their self-preservation at the cost of the global good.

Given the systemic impacts of large-scale AI applications, the community of benefit includes all current and future generations.

Recommendation: Focus on extreme risks and human oversight

The Advisory Body should ensure that regulators comprehend the extreme risks posed by advanced AI systems and the potential tragedy of the commons. The priority for AI governance should focus on maintaining human oversight over these systems, even in the research and development stage.

Risks are particularly high from frontier AI, which currently includes foundation models, given the absence of reliable methods for algorithmic control. In the context of a technology with the potential to influence not just individuals but societies at large, we suggest leveraging the UN's moral authority to emphasize concerns regarding foundation models and underscore the responsibilities of States and companies in reducing global and extinction risks from frontier AI, while also correcting the biases of existing models.

[49] Some Member States have argued that the UN organ's composition, which has not changed since its inception in 1945, is incompatible with geopolitical realities and response.
[50] https://press.un.org/en/2023/sc15263.doc.htm

Step Two: Contextualization

There is no one-size-fits-all approach to AI governance. Various levels of oversight require different governance structures. Pursuing an array of AI governance models across the multilateral ecosystem to accelerate adoption and enable customization seems like the most effective strategy.

The global scope of AI's impacts demands international oversight, but achieving a unified governance regime is unlikely for the foreseeable future. Regulation is primarily implemented and enforced at regional or nation-state levels, especially for rapidly-evolving technology. Thus, contextualizing AI governance for national governments while preserving international coherence is crucial.

Recommendation: Foster policy coherence

The Advisory Body should be focused on driving convergence around best practices and norms in AI governance which can later – once proven effective across contexts – be enshrined in international agreements.

It will be especially important to support governing bodies who are leading regulatory development, such as governments in the UK and EU, and coordinating risk management approaches with US and Chinese AI developers. Meanwhile, regulators should encourage the development of impactful applications of AI technology to accelerate the achievement of the SDGs, especially in Low- and Middle-Income Countries (LMICs).

Step Three: Participatory Decision-making

People will be more likely to follow the rules if they had a hand in writing them.

Given that frontier AI development is concentrated almost entirely in the United States and China, it is important to start with these two countries. The US and China have by far the most leverage to reduce risk. The Advisory Body must engage with and apply pressure to leading AI developers in these countries to work on AI safety and slow development and deployment. The Advisory Body should support the development of a governance infrastructure between these two governments as a global priority.

However, it is also important to ensure that a diverse group of nations and humans participates in any international AI governance process. Democratizing access to the development of frontier AI systems is not advised, however, as it would amplify existing risks.

Recommendation: Secure crucial participants

The Advisory Body should advocate to halt the open sourcing of frontier models until a globally acceptable system for measured risk taking has been developed.

The Advisory Body should also take on the responsibility of convening multinational consultations on the global risks of AI.

The UN is the only legitimate global entity that can meaningfully represent the international community. Leaving oversight solely to a few private US entities is inadequate given AI's global significance. Past UN treaties on shared frontiers like outer space and Antarctica offer precedents for managing AI's global risks. An inclusive international consultation process could examine and determine what the risks and pace of foundation models should be.

On the most ambitious timeline, this consultation process could be announced and endorsed by UN Member States at the upcoming seventy-eighth UN General Assembly in September 2023. The Advisory Body would be tasked with planning and executing this process, including a world tour to convene regional consultations in Asia, Africa, Latin America, and Europe, over the following months. It could then commence a process of synthesizing the findings into a final report. The final step could culminate in a draft Universal Declaration on AI to be announced at the Summit of the Future in 2024.

Step Four: Monitoring

Once rules have been set, verification and compliance processes must be established. The governance of the commons does not rely on good will, but rather accountability.

This form of technical AI regulation could start with tracking the hardware used to train frontier AI models and requiring anyone using large amounts of infrastructure to prove that the models they train meet the highest standards for safety and security. On the software side, developers should employ evaluations to screen models for hazardous traits and incentives. It is important to note that we currently lack consensus on how to effectively perform these evaluations.

This technical regulatory capacity will be developed outside of the UN system, but developing consensus on best practices must be an interdisciplinary and democratic process. Thus, while it is not the UN's role to develop technical expertise, it can be instrumental in institutionalizing and harmonizing risk management solutions when they arise, for example through standards set by the International Organization for Standardization (ISO).

Recommendation: Encourage data-sharing

The Advisory Body should support the implementation of an internationally coherent approach to monitoring, starting with leading AI labs in the US and China, and aiming to develop globally-accepted ISO standards.

This requires helping governments to develop the sufficient regulatory capacity to accurately understand the capabilities of frontier AI systems and perform credible third-party evaluations for their potential misuses and safety/alignment risks. Before this capacity exists, private actors should not be moving ahead with the unrestricted development of frontier AI systems.

Step Five: Sanctions

To enforce oversight, consequences must exist for unsafe AI systems. AI companies need to face legal liability for the harms their systems cause. Liability is one of the few threats that effectively enforces compliance with national regulation among private companies.

The UN can coordinate Member States to impose aligned sanctions on non-compliant companies operating within their borders, such as:

- Fines proportional to damage caused by rogue AI and tied to revenue, increasing until compliance
- Blocking illegal AI services and rescinding government contracts
- Reputational damage through UN-published warnings and restrictions on UN partnerships
- Legal liability for harms caused by AI systems
- Export controls on computing hardware for repeat offenders.

Recommendation: Encourage national adoption of sanctions

The UN is well-positioned to develop guidelines for national legislation, enabling context-appropriate sanctions. Rather than directly enforcing sanctions, the UN can leverage its moral authority and membership to spur the adoption of aligned sanction regimes nationally. Through model legislation and multilateral pressure, consequences for irresponsible AI can be institutionalized across jurisdictions. The threat of coordinated exclusion from major markets provides deterrence.

In summary, the UN should develop sanction guidelines, encourage national adoption, and coordinate restrictions on rogue actors' operations. Aligned deterrents instituted nationally, guided globally, offer a decentralized enforcement model suited to AI.

Step Six: Conflict Resolution

When encountering difficulties in the enforcement of regulations, addressing these challenges should follow an uncomplicated, economical, and approachable process, with appropriate channels to vocalize complaints and seek guidance. Mediation requires an impartial third-party actor to make respected judgements. In AI governance, this could include deciding whether something is a frontier AI model; whether a nation is violating previous agreements; or whether an AI company is ignoring its safety obligations.

Recommendation: Develop mediation capacity

Leveraging the UN's experience with conflict resolution and intercultural communication, the UN could develop a key role in developing diplomatic capacity for private companies. The United Nations Commission on International Trade Law (UNCITRAL) has established rules for

arbitration that are widely used in resolving international commercial disputes. This might allow labs to coordinate across borders more closely in the near future while in the medium-term, at national and international levels, building up the technical understanding required to develop new governing bodies. In the long run, one should aim for an arbitration instance to authoritatively resolve conflicts. However, the current pace of AI development demands a less formal and swifter way for parties to resolve disputes.

Step Seven: Legitimacy

For oversight to be effective, governing bodies must have buy-in from key actors who view them as legitimate authorities. This requires inclusion of private sector developers in technical governance, while averting risks of regulatory capture.

Private companies possess critical insights needed to craft pragmatic policies, not just self-interested ones. However, profit motives may not fully align with societal priorities. Still, substantial corporate participation is crucial, coupled with accountability norms.

The UN can enable constructive exchange through multi-stakeholder fora, building mutual understanding between developers, government, and civil society regarding capabilities and risks. Transparent engagement demonstrates that innovation and oversight can co-exist.

Recommendation: Facilitate public-private coordination

- Develop guidelines for accountable industry participation in governance bodies
- Convene inclusive spaces for developers, regulators, and the public to collaborate
- Encourage proactive corporate transparency and cooperation on risks
- Provide policy guardrails so oversight evolves collaboratively, not reactively
- Blend technical insights with public values and oversight through joint fora.

If cultivated properly, public-private coordination will prove more fruitful than unilateralism in governing globally impactful technologies like AI. Blending capabilities with oversight and values can enable broadly accepted, legitimate governance.

Step Eight: Networked Multilateralism

Global issues require a multi-tiered approach. Regulation need not be centrally organized to be globally respected, as long as it develops in convergent ways in different places. A decentralized, yet coordinated approach, seems plausible as all governments are concerned about the potential threats from automated 'black box' decision-making systems.

Recommendation: Build engagement capacity

The Advisory Body should engage in various international fora and utilize regional and 'minilateral' partnerships as proving grounds for governance models before scaling successful

versions globally. It can also catalyse global AI governance by cultivating and supporting national champions to lead by domestic example and advocate internationally.

The Advisory Body can also ensure the inclusion of under-resourced actors representing crucial populations. Particularly low- and middle-income countries should be direct beneficiaries of technological progress and protected from harms by strengthening their infrastructure and governance capacity. A first step could be the establishment of an international training system to rapidly upskill governments and encourage the development of high-impact AI applications in LMICs.

Summary of Recommendations

Step	Recommendation	Description
1	Focus on human agency	Focus regulatory attention on the risk of losing human agency, starting at the research and development stage of frontier AI.
2	Foster policy coherence	Foster policy transfer for effective risk management in frontier AI while encouraging the development of empowering applications.
3	Secure crucial participants	Secure the participation of the USA and China in the development of an international AI governance regime.
4	Encourage data-sharing	Encourage data sharing for the development of international AI development monitoring infrastructure.
5	Support national legislation	Support the development of hard law at regional and national levels to ensure technological development proceeds at a manageable pace.
6	Develop mediation capacity	Develop mediation capacity to foster high-trust spaces for coordination between leading AI developers and regulators.
7	Facilitate public-private coordination	Boost multi stakeholder fora for knowledge exchange and public accountability between frontier AI developers, regulators, and the global community.
8	Build engagement capacity	Build engagement capacity for low- and middle-income countries to participate to fairly shape global benefit distribution.

Conclusion

The long-term governance of foundation AI models may eventually lead to a global AI regulatory framework signed by all 193 UN Member States with stringent safety requirements and an effective auditing mechanism. As the above model analyses show, there is significant value in granting enforcement authority to such an institution.

However, binding international AI governance will require years of capacity and trust-building. Meanwhile, the short term requires a response that is both relevant and can help lay the groundwork for eventually establishing an international governance regime for advanced AI. We believe the establishment of a UN Advisory Board which is inclusive of Global South voices, yields norm-building and defines principles, supports the pooling of expertise and domestic regulation, and initiates a multistakeholder process is the best mechanism to do so.

Acknowledgements

We are grateful to the following people for discussion and input: Yoshua Bengio, Konrad Seifert, Adam Day, Lennart Heim, Jakob Graabak, Dexter Docherty, Maxime Stauffer, Robert Trager, and Eleonore Fournier-Tombs.