09.09.2023

Global Governance of High-Risk Artificial Intelligence

Stephan Llerena[1]

Abstract

The pursuit of artificial general intelligence (AGI) poses existential risks to humanity. Global cooperation is needed to mitigate these risks. This Article is among the first to propose objectives and ideal features of a global governance regime focused on AGI and other high-risk artificial intelligence (AI). It utilizes a risk-based approach to suggest mechanisms and legal frameworks for addressing two objectives: safety and accountability. Along with noting ideal characteristics of such a global AI safety regime, this Article explores precedent in existing, comparable international systems to suggest possible components of and entities for global governance of high-risk AI.

[1] Summer Research Fellow, Existential Risk Laboratory, University of Chicago. J.D., *magna cum laude*, University of Michigan Law School. I am deeply grateful to Dr. José Jaime Villalobos for invaluable feedback and guidance. For insightful comments and insights, I also owe special thanks to Julian Baldwin, Allison Huang, Yuxin Ji, Miles Wang, and Shuyuan Wang. All views are my own.

Introduction

Humanity is charging toward the unknown. Leading artificial intelligence (AI)[2] companies–often called AI labs–appear to be in a race to develop artificial general intelligence (AGI).[3] While used with varying definitions, AGI is generally understood to be an AI that "has the potential to perform a wide range of tasks and exhibit cognitive abilities comparable to or surpassing human intelligence."[4] Although some argue that different states are in an AGI race as well, the United States and its allies currently have a dominant position in AGI development.[5] The possible benefits of AGI are almost innumerable–AGI would drive new medical research and forms of treatment; help optimize the use of energy and other scarce resources; and improve production processes in many industries.[6]

In pursuing AGI, many AI labs have already created highly capable, "narrow" AIs that are exceptional in specific tasks and that are benefiting humanity. For example, Google DeepMind created AlphaFold, a model that "can accurately predict 3D models of protein structures" from sequences of amino acids, addressing a decades-old problem that will enable researchers to more quickly find new medicines, understand diseases, and pursue other scientific breakthroughs.[7] Google DeepMind also widely shares AlphaFold's protein structure predictions with members of the scientific community, demonstrating that developers of highly capable AIs may aim to widely share the benefits of their models.[8]

However, despite the possible benefits of AI and the positive intentions of many AI developers, the pursuit of AGI poses enormous risks to humanity. Many leading AI researchers

---

[2] "Artificial intelligence" may be defined as a "system that is designed to operate with elements of autonomy and that, based on machine and/or human-provided data and inputs, infers how to achieve a given set of objectives using machine learning and/or logic- and knowledge-based approaches, and produces system-generated outputs such as content (generative AI systems), predictions, recommendations or decisions, influencing the environments with which the AI system interacts." Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts, https://artificialintelligenceact.eu/wp-content/uploads/2022/11/AIA-CZ-Draft-for-Coreper-3-Nov-22.pdf [hereinafter EU AI Act, Nov. 2022 Draft].

[3] *See* Dylan Matthews, *The $1 billion gamble to ensure AI doesn't destroy humanity*, Vox, July 17, 2023, https://www.vox.com/future-perfect/23794855/anthropic-ai-openai-claude-2 (describing how Anthropic, a leading AI lab, seeks to be "safety-first," while also developing AGI); Google DeepMind, https://www.deepmind.com/about ("Our long term aim is to solve intelligence, developing more general and capable problem-solving systems, known as artificial general intelligence (AGI)"); Sam Altman, *Planning for AGI and beyond*, OpenAI, Feb. 24, 2023, https://openai.com/blog/planning-for-agi-and-beyond ("Our mission is to ensure that artificial general intelligence–AI systems that are generally smarter than humans–benefits all of humanity").

[4] XenonStack, *Rise of Artificial Intelligence (AGI)*, https://www.linkedin.com/pulse/rise-artificial-general-intelligence-agi-xenonstack/.

[5] Haydn Belfield & Christian Ruhl, *Why policy makers should beware claims of new 'arms races,'* Bulletin of the Atomic Scientists, July 14, 2022, https://thebulletin.org/2022/07/why-policy-makers-should-beware-claims-of-new-arms-races/#post-heading (arguing false assumptions of "arms races" needlessly hasten the development of dangerous technology).

[6] *Supra*, n. 4.

[7] *AlphaFold*, Google DeepMind, https://www.deepmind.com/research/highlighted-research/alphafold.

[8] *See id.*

have argued that AGI itself would pose an existential risk to humanity.[9] An "existential risk" may be defined as a risk "that threaten[s] the destruction of humanity's long-term potential," including outcomes such as extinction.[10] Hendrycks, Mazeika, and Woodside further argue that AIs are subject to various sources of extreme risk, including from malicious use for purposes, such as novel forms of bioterrorism; overreliance on flawed AIs in autonomous warfare, including potential nuclear conflicts; organizational risks; and rogue AIs whose actions do not align with human preferences.[11]

Given the global risks posed by AGI and other high-risk AI, many AI researchers, industry leaders, and policymakers argue that global cooperation is needed.[12] To the extent that states or AI labs in different states are racing or will race to create AGI for military or civilian purposes, international cooperation imposing uniform protective safety measures would facilitate wider adoption of safety practices, as actors would know their competitors are more likely to comply with such standards.[13] Further, if a rogue AGI is truly an existential risk to humanity, it does not matter which state or AI lab first creates it–preventing the arrival of such a rogue AGI is in every state's interest.[14]

States and intergovernmental organizations are beginning to signal an openness to international cooperation, but such initiatives are in early stages. For example, as of August 2023, the United States has advocated for largely non-binding forms of international cooperation.[15] The United Kingdom has expressed its willingness to lead on international

---

[9] Richard Ngo, *AGI Safety from First Principles*, Sep. 2020, https://drive.google.com/file/d/1uK7NhdSKprQKZnRjU58X7NLA1auXlWHt/view; Dan Hendrycks & Mantas Mazeika, *X-Risk Analysis for AI Research*, Sep. 20, 2022, https://arxiv.org/pdf/2206.05862.pdf [hereinafter Hendrycks & Mazeika, *X-Risk Analysis*]; Eliezer Yudkowsky, *Pausing AI Developments Isn't Enough. We Need to Shut it All Down*, Time, March 29, 2023, https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough/?ref=campaignforaisafety.org; Benjamin S. Bucknall & Shiri Dori-Hacohen, *Current and Near-Term AI as a Potential Existential Risk Factor*, Sep. 21, 2022, https://arxiv.org/pdf/2209.10604.pdf; Dan Hendrycks, Mantas Mazeika, & Thomas Woodside, *An Overview of Catastrophic AI Risks*, July 11, 2023, https://arxiv.org/pdf/2306.12001.pdf; Yoshua Bengio, *FAQ on Catastrophic AI Risks*, June 24, 2023, https://yoshuabengio.org/2023/06/24/faq-on-catastrophic-ai-risks/.

[10] Toby Ord, The Precipice 6 (2020). In line with Ord's definition, Martínez and Winter confirmed that many experts interpret "existential risk" to only refer to risks that would endanger virtually all of humanity, while lay persons considered certain risks endangering a lower minimum amount of lives to be existential. Eric Martínez & Christoph Winter, *Ordinary Meaning of Existential Risk* (Legal Priorities Project Working Paper No. 7-2022), at 23, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4304670.

[11] Dan Hendrycks, Mantas Mazeika, & Thomas Woodside, *An Overview of Catastrophic AI Risks*, July 11, 2023, at 2, https://arxiv.org/pdf/2306.12001.pdf.

[12] Yonadav Shavit, *What Does it Take to Catch a Chinchilla? Verifying Rules on Large-Scale Neural Network Training via Compute Monitoring*, https://arxiv.org/pdf/2303.11341.pdf [hereinafter Shavit, *Compute Monitoring*]; Michelle Toh & Yoonjung Seo, *OpenAI CEO calls for global cooperation to regulate AI*, June 9, 2023, CNN Business, https://www.cnn.com/2023/06/09/tech/korea-altman-chatgpt-ai-regulation-intl-hnk/index.html.

[13] Allan Dafoe, *AI Governance: A Research Agenda*, Aug. 27, 2018, at 45–6, https://www.fhi.ox.ac.uk/wp-content/uploads/GovAI-Agenda.pdf.

[14] *See* TED, *Will Superintelligent AI End the World? | Eliezer Yudkowsky*, YouTube, July 12, 2023, https://www.youtube.com/watch?v=Yd0yQ9yxSYY (arguing for international bans on certain AGI development).

[15] *See* National Security Commission on Artificial Intelligence, *Chapter 15: A Favorable International Technology Order*, https://reports.nscai.gov/final-report/chapter-15 (advocating for collaboration with U.S. allies to develop AI principles, standards, and expertise, as well as investment); United Nations, *International Community Must Urgently Confront New Reality of Generative, Artificial Intelligence, Speakers Stress as Security Council Debates Risks,*

agreements for AI regulation, but its government initiatives are still in an exploratory phase.[16] While not expressly stating it will advocate for global cooperation, China has recognized the existential threat of AI.[17] At the supranational level, the Council of Europe has tasked a Committee on Artificial Intelligence (CAI) to draft a legally-binding, global instrument: the framework Convention on Artificial Intelligence, Human Rights, Democracy, and the Rule of Law.[18] Also, both the United Nations Secretary-General[19] and the Security Council[20] have initiated efforts for global cooperation. Further, with a Council, Steering Committee, Secretariat hosted by the Organization for Economic Co-operation and Development (OECD), and two centers of research expertise, the Global Partnership on Artificial Intelligence (GPAI) acts as a "multistakeholder initiative bringing together leading experts . . . for sharing multidisciplinary research and identifying key issues among AI practitioners."[21]

Similarly, scholarship on proposals for global governance regimes of high-risk AI is in early stages. While researchers and commentators have proposed general structures[22] or specific mechanisms[23] useful to possible AGI global governance organizations, many of these proposals either do not explain the functioning of possible organs of such organizations, or they do not structure the proposals to account for the unique characteristics of governing AI. For instance, in response to proposals calling for an AGI organization similar to the International Atomic Energy Agency (IAEA), nuclear experts have cautioned about the differences between the two technologies.[24] More research is needed to explore ideal structures for different possible functions and organs of a global AGI governance body, considering a variety of important objectives.

This Article builds on the existing literature to propose a tailored, new global governance system for certain existential risks from high-risk AI. It is among the first articles to link a global

---

Rewards, July 18, 2023, https://press.un.org/en/2023/sc15359.doc.htm [hereinafter UN 9381st Meeting] (noting the United States in early 2023 "proposed a political declaration on the responsible military use of AI").

[16] Tech entrepreneur Ian Hogarth to lead UK's AI Foundation Model Taskforce, June 18, 2023, Gov.uk, https://www.gov.uk/government/news/tech-entrepreneur-ian-hogarth-to-lead-uks-ai-foundation-model-taskforce; UN 9381st Meeting, *supra* note 13 (calling for global governance of AI consistent with certain democratic principles).

[17] Lauren Sforza, *China warns of 'complicated and challenging circumstances' posed by AI risk* (May 31, 2023), The Hill, https://thehill.com/policy/technology/4028141-china-warns-of-complicated-and-challenging-circumstances-posed-by-ai-risk/.

[18] Council of Europe, *CAI - Committee on Artificial Intelligence*, https://www.coe.int/en/web/artificial-intelligence/cai#%7B%22126720142%22:[]%7D.

[19] António Guterres, *Secretary-General's remarks to the Security Council on Artificial Intelligence*, July 18, 2023, https://www.un.org/sg/en/content/sg/speeches/2023-07-18/secretary-generals-remarks-the-security-council-artificial-intelligence ("I am convening a multistakeholder High-Level Advisory Board for Artificial Intelligence that will report back on the options for global AI governance, by the end of this year.").

[20] UN 9381st Meeting, *supra* note 13.

[21] The Global Partnership on Artificial Intelligence, About GPAI, https://gpai.ai/about/.

[22] *See, e.g.*, Lewis Ho et al., *International Institutions for Advanced AI*, July 11, 2023, https://arxiv.org/pdf/2307.04699.pdf [hereinafter Ho, *International Institutions*].

[23] *See e.g.*, Shavit, *Compute Monitoring*, *supra* note 10.

[24] *See, e.g.*, Ian J. Stewart, *Why the IAEA model may not be best for regulating artificial intelligence*, June 9, 2023, Bulletin of the Atomic Scientists, https://thebulletin.org/2023/06/why-the-iaea-model-may-not-be-best-for-regulating-artificial-intelligence/.

AI governance regime's objectives to existential risks from AI, specifying possible entities and mechanisms to pursue objectives relating to high-risk AI. Given this approach, it recognizes but does not focus on the urgent need for dedicated global cooperation on AI safety research.[25] Further, this Article recognizes but does not cover the additional possible need for global governance of non-existential risks from AI.[26] Rather, it envisions its suggestions as being capable of complementing or providing a structure for other global AI governance initiatives.

To address existential risks from high-risk AI, this Article focuses on entities and mechanisms that would fulfill two objectives: maintaining and enforcing protective safety measures for high-risk AI; and enabling accountability of major actors working with and toward high-risk AI.

Part I provides an in-depth background of high-risk AI, including the process through which an AI is trained. Further, it argues that a tiered, risk-based approach to global governance of high-risk AI is optimal, as such an approach would ensure that the regime is not needlessly intrusive to minor AI applications and is clear to the subjects of regulation.

Part II provides an overview of an existential risk that creates the need for global AI protective safety measures, as well as different ways to approach future threats. It further explores initial possible safety measures, as well as ideal characteristics of legal structures that would supervise the implementation, maintenance, and enforcement of such safety measures. Based on existing structures, it concludes by proposing possible legal entities for global governance of high-risk AI.

Part III covers accountability mechanisms for high-risk AI. It details relevant existential risks from AI misuse that necessitate accountability mechanisms and then examines precedents and possible applications for transparency, oversight, complaint, and enforcement mechanisms to guard against such misuse.

The Article concludes by summarizing its contributions to the existing literature, acknowledging challenges to global cooperation for high-risk AI governance and suggesting areas for further scholarship.

I. A Background on AI and Risk-Based Scopes for Governance

To provide context for possible points of regulation, this Part provides an overview to the process of developing powerful AIs that may pose existential risks. It covers deep learning,

---

[25] For such proposals, see Ho, *International Institutions*, *supra* note 20; Daniel Zhang et al., *Enhancing International Cooperation in AI Research: The Case for a Multilateral AI Research Institute*, Stanford University Human-Centered Artificial Intelligence, https://hai.stanford.edu/sites/default/files/2022-05/HAI%20Policy%20White%20Paper%20-%20Enhancing%20International%20Cooperation%20in%20AI%20Research.pdf; Sophie-Charlotte Fischer & Andreas Wenger, *A Politically Neutral Hub for Basic AI Research*, March 2019, https://css.ethz.ch/content/dam/ethz/special-interest/gess/cis/center-for-securities-studies/pdfs/PP7-2_2019-E.pdf.

[26] For a non-exhaustive list of such risks, see Bernard Marr, *The 15 Biggest Risks of Artificial Intelligence*, June 2, 2023, Forbes, https://www.forbes.com/sites/bernardmarr/2023/06/02/the-15-biggest-risks-of-artificial-intelligence/?sh=7cf25eb82706.

foundation models, the training process for such models, and the hardware involved in these processes. Given the commonly-held belief that large language models (LLMs) are the AIs most likely to become AGIs,[27] it devotes particular focus to the training process for such foundation models. Further, it discusses the difficulty of governance based on definitions of AI and advocates for a tiered, risk-based approach for global governance.

A. Training a Foundation Model

Drawing inspiration from the human brain, neural networks are at the heart of deep learning, which currently drives AI development. Deep learning (DL) refers to a subset of machine learning (ML)[28] through which neural networks, or processing nodes organized into deep layers, process large amounts of data to provide outputs based on identified probabilistic patterns.[29] Neural networks can be further organized into different ML architectures, which impact an AI's efficiency in converting inputs to desired outputs; currently, the dominant ML architecture for text-generating AI is known as the transformer.[30]

Like neurons in the human brain, nodes "fire" if inputs provided by a previous layer surpass a threshold value defined in an activation function, resulting in activations for the next layer of nodes.[31] Each node's activation function is influenced by weights assigned to each of the connections between a given node and the nodes in the previous layer, as well as a "bias" that impacts the threshold needed to be surpassed before a node activates.[32]

Thus, "training" a powerful AI involves adjusting the parameters of an AI's connections between its nodes to produce desired outputs. Parameters are not intentionally or selectively adjusted by humans; rather, they are adjusted through optimization methods, such as stochastic gradient descent.[33] While making such adjustments eventually may produce desired outputs, the functioning–and by extension control–of neural networks is poorly understood in part because of the human inability to account for parameters' effect on a given output, an issue referred to as the "black box problem."[34]

Further, training powerful AI systems, including foundation models, to adjust their parameters to produce desired outputs is a resource-intensive process. A foundation model is a

---

[27] *See* Sébastien Bubeck et al., *Sparks of Artificial General Intelligence: Early experiments with GPT-4*, April 13, 2023, Microsoft Research, https://arxiv.org/pdf/2303.12712.pdf [hereinafter, Bubeck, *Sparks of AGI*].
[28] Machine learning refers to "an application of artificial intelligence that is characterized by providing systems the ability to automatically learn and improve on the basis of data or experience, without being explicitly programmed." 15 U.S.C. § 9401(11) (2021).
[29] Laurie A. Harris, Cong. Rsch. Serv., R46795, Artificial Intelligence: Background, Selected Issues, and Policy Considerations, at 4 (2021).
[30] Ashish Vaswani et al., *Attention is All You Need*, Aug. 2, 2023, Google, https://arxiv.org/pdf/1706.03762.pdf.
[31] 3Blue1Brown, *But what is a neural network? | Chapter 1, Deep learning*, YouTube, Oct. 5, 2017, https://www.youtube.com/watch?v=aircAruvnKk.
[32] Together, the weights and biases of an AI system constitute its "parameters." *Id.*
[33] Aishwarya V. Srinivasan, *Stochastic Gradient Descent–Clearly Explained*, Sept. 7, 2019, https://towardsdatascience.com/stochastic-gradient-descent-clearly-explained-53d239905d31.
[34] *AI's mysterious 'black box' problem, explained* (Mar. 6, 2023), University of Michigan-Dearborn,https://umdearborn.edu/news/ais-mysterious-black-box-problem-explained.

"model that is trained on broad data . . . that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks . . . ."[35] For instance, the systems behind OpenAI's ChatGPT (GPT-3, GPT-3.5, and GPT-4) are foundation models with billions of parameters that produce conversational text as the desired outputs.[36] Text-generating foundation models are also referred to as large language models (LLMs), but the term "foundation model" is more inclusive of powerful models, as many models are "multimodal (e.g. possess visual capabilities)."[37]

Generally, the training process of a foundation model may consist of two steps: (1) generative pretraining and (2) fine-tuning, which may be accomplished in different manners but commonly consists of (i) supervised fine-tuning and (ii-a) reinforcement learning from human feedback (RLHF) or (ii-b) a process known as constitutional AI.[38] In each step, the model is fine-tuned from the previous step's resulting model, resulting in changes to its parameters.[39]

First, in generative pretraining, the model undertakes unsupervised learning[40] based on massive amounts of text from the internet, resulting in a raw model with parameters that produce probabilistic responses to inputs based on the text used in pre-training.[41] This process requires the "concurrent use of . . . thousands of specialized accelerators with high inter-chip communication bandwidth" ("ML chips") to conduct needed calculations for months at a time, as well as large amounts of computing power ("compute").[42] The overwhelming majority of companies producing high-quality ML chips are based in the United States or with its allied states, including Google, Nvidia Corporation, and Advanced Micro Devices, Inc. (AMD).[43] Manufacturing of advanced ML chips is overwhelmingly dominated by Taiwan Semiconductor Manufacturing Corporation (TSMC),[44] and the Dutch company ASML Holdings has a

---

[35] Rishi Bommasani et al., *On the Opportunities and Risks of Foundation Models*, at 3, https://arxiv.org/pdf/2108.07258.pdf.

[36] *Id.* As of August 2023, ChatGPT, a conversational AI capable of producing human-like text for a wide range of situations, was the most rapidly adopted web application in history. Cindy Gordon, *ChatGPT is the Fastest Growing App in the History of Web Applications*, Forbes, Feb. 2, 2023, https://www.forbes.com/sites/cindygordon/2023/02/02/chatgpt-is-the-fastest-growing-ap-in-the-history-of-web-applications/?sh=1e2b1efe678c.

[37] Markus Anderljung et al., *Frontier AI Regulation: Managing Emerging Risks to Public Safety*, July 11, 2023, at 7, n. 8 [hereinafter, Anderjlung et al., *Frontier AI Regulation*].

[38] Ari Seff, How ChatGPT is Trained, YouTube, Jan. 24, 2023, https://www.youtube.com/watch?v=VPRSBzXzavo; Yuntao Bai et al., *Constitutional AI: Harmlessness from AI Feedback* (Dec. 15, 2022), Anthropic, https://arxiv.org/pdf/2212.08073.pdf. See *infra* for explanations of these processes.

[39] Ari Seff, How ChatGPT is Trained, YouTube, Jan. 24, 2023, https://www.youtube.com/watch?v=VPRSBzXzavo.

[40] Unsupervised learning uses ML algorithms to analyze and cluster datasets, finding patterns in data without human intervention, while supervised learning uses labeled data. *What is unsupervised learning?*, IBM, https://www.ibm.com/topics/unsupervised-learning#:~:text=Unsupervised%20learning%2C%20also%20known%20as,the%20need%20for%20human%20intervention.

[41] Ari Seff, How ChatGPT is Trained, YouTube, Jan. 24, 2023, https://www.youtube.com/watch?v=VPRSBzXzavo.

[42] Shavit, *Compute Monitoring*, *supra* note 10, at 1.

[43] *See id*. Notably, developing "the same AI application using older AI chips or general-purpose chips can cost tens to thousands of times more." Saif M. Khan & Alexander Mann, *AI Chips: What They Are and Why They Matter: An AI Chips Reference* (April 2020), at 3, Center for Security and Emerging Technology, https://cset.georgetown.edu/publication/ai-chips-what-they-are-and-why-they-matter/.

[44] *Taiwan's dominance of the chip industry makes it more important*, The Economist, Mar. 6, 2023, https://www.economist.com/special-report/2023/03/06/taiwans-dominance-of-the-chip-industry-makes-it-more-important.

near-monopoly on advanced lithography machines, which are needed to produce cutting-edge ML chips.[45]

Despite the high amounts of compute, capital, and high-quality ML chips needed for the generative pretraining of a quality foundation model, various advancements are reducing these barriers.[46] For instance, Zeus, an energy optimization framework developed at the University of Michigan, may reduce compute by up to 75% for the development of a given foundation model, while maintaining the same hardware and datacenter infrastructure.[47] Further, advancements in ML chips, processors, systems, and algorithms facilitate much more efficient development of foundation models, reducing the time, compute, and capital needed to produce a given foundation model.[48] Thus, given these technological advancements, generative pretraining for a foundation model is expected to become cheaper and more accessible as time progresses.

Second, after generative pretraining results in a raw foundation model, fine-tuning is used to train the model to produce text that mimics that produced by a human, as well as to reduce the instances where the model provides harmful or offensive information.[49] In the first step of fine-tuning, supervised fine-tuning is applied to the model; in this process, the model is trained by data from tens of thousands of conversations generated by humans, resulting in a model that is better at engaging in a question-and-answer exchange than the previous raw model.[50]

In the second step of fine-tuning, RLHF or constitutional AI may be applied.[51] For RLHF, thousands of people examine the model's chat outputs, ranking them "according to criteria, such as appropriateness, accuracy, politeness, and avoidance of improper topics."[52] From these data produced from human feedback, the AI learns to improve its outputs, reducing–but not eliminating–the risk of outputs that would provide harmful information, such as guidance for developing bioweapons.[53] While providing an effective manner of fine-tuning, RLHF is

---

[45] Xinmei Shen, *Chinese imports of ASML lithography chip-making machines have surged past the Dutch company's 2023 estimates*, South China Morning Post, Aug. 26, 2023, https://www.scmp.com/tech/tech-war/article/3232401/chinese-imports-asml-lithography-chip-making-machines-have-surged-past-dutch-companys-2023-estimates.

[46] "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, (3) the effectiveness of the algorithms used to train it. The quantity of chips used to train a model is increasing by 2x-5x per year. Speed of chips is increasing by 2x every 1-2 years. And algorithmic efficiency is increasing by roughly 2x per year." Written Testimony of Dario Amodei, For a hearing on "Oversight of A.I.: Principles for Regulation," July 25, 2023, Before the Judiciary Committee, https://www.judiciary.senate.gov/imo/media/doc/2023-07-26_-_testimony_-_amodei.pdf [hereinafter, Amodei Testimony].

[47] Zachary Champion, *Optimization could cut the carbon footprint of AI training by up to 75%*, April 17, 2023, University of Michigan, https://news.umich.edu/optimization-could-cut-the-carbon-footprint-of-ai-training-by-up-to-75/.

[48] *Supra*, note 46.

[49] Ari Seff, How ChatGPT is Trained, YouTube (Jan. 24, 2023), https://www.youtube.com/watch?v=VPRSBzXzavo.

[50] Written Testimony of Stuart Russell, For a hearing on "Oversight of A.I.: Principles for Regulation" (July 25, 2023), Before the Judiciary Committee, https://www.judiciary.senate.gov/imo/media/doc/2023-07-26_-_testimony_-_russell.pdf [hereinafter, Russell Testimony].

[51] *Id.*

[52] *Id.*

[53] *Id.*

imperfect and has significant problems, including relying on manipulable, fallible human evaluators and possibly pushing the raw model in dangerous directions before resulting in a "safer" model.[54]

Instead of using RLHF, an AI lab may subsequently follow supervised fine-tuning with a process known as constitutional AI. With this approach, a model "itself ranks and critiques its own possible outputs based on a set of principles, stated in English, concerning behaviors that are allowable . . . but there is no guarantee that the machine-generated rankings are comparable to human feedback."[55]

After it is fine-tuned, an AI may reach the public in different manners. Commonly, an AI is offered to the public or customers through an application programming interface (API) [or chat user interface], which essentially is an intermediary between two applications that may also integrate safety filters to mitigate errors or reduce risks from an AI.[56] Alternatively, the developer of an AI may open source the model, releasing its parameters and relevant code, enabling the public to replicate that model or further fine-tune it.[57] While many AI developers have become increasingly hesitant to open source their models, Meta appears committed to open sourcing its foundation models.[58]

B. A Risk-Based Scope for Global AI Governance

This section discusses possible approaches for defining the scope of global regulation of certain existential risks from AI, determining that a risk-based approach is the best option. In other words, this section seeks to answer the question of *what* AI would be regulated under this Article's proposed governance regime for existential risks from AI. In an assessment of ways to determine scope, or which AIs should be regulated, Jonas Schuett argues policymakers should "only use the term *AI* for the scope definition if there is a good definition of *AI*."[59] Similarly, I argue that this Article's proposed global governance regime should only rely on a term to determine its scope of regulation if there is a good legal definition for AIs that pose existential risks.

Based on the legal traditions of the United States and the European Union, Schuett identifies requirements for legal definitions, arguing they should be used to determine if a

---

[54] Charbel-Raphaël, *Compendium of problems with RLHF* (Jan. 29, 2023), https://www.lesswrong.com/posts/d6DvuCKH5bSoT62DB/compendium-of-problems-with-rlhf.

[55] Russell Testimony, *supra*, at footnote 14.

[56] Rem Darbinyan, *What are AI APIs, and How Do they Work?* (April 13, 2022), https://www.dataversity.net/what-are-ai-apis-and-how-do-they-work/. For example, ChatGPT is accessible to users through an API, but if the underlying parameters of the foundation model are not public, it would be near-impossible for a third party to further fine-tune or recreate the underlying model itself.

[57] *What is an Open Source Model?*, Iguazio, https://www.iguazio.com/glossary/open-source-model/#:~:text=Open%20source%20has%20always%20been,of%20its%20open%2Dsource%20license.

[58] *See* Melissa Heikkilä, *Meta's latest AI model is free for all*, MIT Technology Review, https://www.technologyreview.com/2023/07/18/1076479/metas-latest-ai-model-is-free-for-all/.

[59] Jonas Schuett, *Defining the scope of AI regulations*, 15 Law, Innovation and Technology __, 4 (forthcoming) (emphasis added).

term-based approach is ideal for setting a governance regime's scope.[60] Schuett specifies six requirements, stating legal definitions: (1) must not be over-inclusive, which would include cases that are not in need of regulation based on the regulation's objective; (2) must not be under-inclusive; (3) must be precise, enabling a determination of whether a "particular case falls under the definition"; (4) must be understandable–ideally based on words' ordinary meaning; (5) should be practical, including clear legal elements; and (6) should be flexible, accommodating technical progress.[61]

Thus, it is difficult to draft a legal definition for AI, especially for those that may pose existential risks. For example, the EU AI Act draft from November 2022 offers a carefully drafted definition for an "artificial intelligence system," reflecting a choice to move away from an earlier definition of AI that largely delegated the definition to software developed through methods listed in an annex.[62] However, while foundation models are currently thought to be the means through which AGI may be achieved, if at all, even a legal definition for such models that satisfies Schuett's requirements is difficult to draft. For instance, although the term is not completely synonymous with foundation models, the November 2022 draft of the EU AI Act does define "general purpose AI system":

> an AI system that–irrespective of how it is placed on the market or put into service, including as open source software–is intended by the provider to perform generally applicable functions such as image and speech recognition, audio and video generation, pattern detection, question answering, translation and others; a general purpose AI system may be used in a plurality of contexts and be integrated in a plurality of other AI systems;[63]

Using Schuett's requirements to evaluate the above definition's effectiveness for capturing AIs capable of posing existential risks–or even solely AIs capable of becoming AGI–it would be difficult to establish the scope of regulation based on this term. The largest problem would be over-inclusiveness. The above definition would include past foundation models, such as GPT-3, which definitely does not pose an existential threat.[64] Further, even the definition recognizes the "plurality of contexts" where it would be applicable.[65] Given finite resources of regulatory bodies and possible resistance from entities whose models are needlessly regulated, over-inclusiveness weighs heavily against using the above term to define the scope of regulation. It is true the definition is not greatly under-inclusive, and it is precise, understandable, and

[60] *Id.* at 6–9.
[61] *Id.* at 7.
[62] Compare EU AI Act, Nov. 2022 Draft definition of AI system, *supra*, note 2, with art. 3(1) of April 2021 Draft of EU AI Act, https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF ("software . . . developed with one or more of the techniques and approaches listed in Annex I . . .") [hereinafter EU AI Act, April 2021 Draft].
[63] EU AI Act, Nov. 2022 Draft, Article 3(1b).
[64] **[Cite source about GPT-3's competencies]**
[65] See footnote 53.

flexible. However, it would be difficult to base a global governance regime on such an over-inclusive definition.

Relative to a scope based on defined terms, a risk-based approach could perform comparably well with Schuett's requirements, and it would enable a more practical global regulatory system. Risk-based regulation may be defined as the use of "decision-making frameworks and procedures to prioritize regulatory activities and the deployment of resources according to an assessment of the risks . . . [relative to] objectives."[66] This approach often results in tiers of regulation, applying higher scrutiny or restrictions to riskier activities.

To help define these tiers of risk, Schuett offers categories for determining main sources of risk from AI, including "(1) technical approaches ('how it is made'), (2) applications ('what it is used for'), and (3) capabilities ('what it can do')."[67] He also notes that definitions can better meet his stated legal requirements through the use of exemptions to reduce over-inclusiveness; catch-all definitions for future flexibility; and sunset clauses or built-in revision schedules to enable updates to scope definitions.[68] For certain definitions drafted using a combination of technical approaches, applications, capabilities, and the aforementioned tools, regulators can better meet the requirements for legal definitions to determine regimes' scope of regulation, instead of relying on a singular term's definition.[69]

Notably, despite its increasingly precise definitions for AI, the EU AI Act continues to largely rely on a risk-based approach to frame its governance system, estimating risk based on threats to "public interests, such as health and safety and the protection of fundamental rights, as recognised and protected by Union law."[70] Based on these risks, the EU AI Act divides AIs into tiers of models that are (1) prohibited, (2) high-risk, (3) subject to transparency obligations, or (4) subject to no restrictions but encouraged to comply with future voluntary codes of conduct.[71] To determine which AIs are prohibited or deemed high-risk, the Draft EU AI Act designates certain categories of AIs; for instance, high-risk AIs include systems subject to any of 12 product safety regulatory regimes or falling under any of eight categories.[72]

However, as with all tiered, risk-based regimes, how much regulatory scrutiny applies to a system depends on the classification system itself–a significant disadvantage for flexibility for future technological advancements. While the EU AI Act does not currently facilitate additions to listed prohibited AIs–apart from amendments to the Act itself–it empowers the European Commission to amend the list of high-risk AIs, as long as two conditions are fulfilled: (1) the AI

---

[66] Robert Baldwin & Julia Black, *Driving Priorities in Risk-based Regulation: What's the Problem?*, 43 Journal of Law and Society 565, 567 (2016).
[67] Schuett, *supra* note 49, at 18.
[68] *Id.* at 26, 29–30.
[69] *Id.* at 24–25.
[70] EU AI Act, Nov. 2022 Draft, Annex (5), at pg. 5.
[71] *See id.* at Titles II, III, IV, & IX.
[72] For a concise summary of the AI Act's tiers, scope, requirements, and sanctions, see Charlotte Siegmann & Markus Anderljung, *The Brussels Effect and Artificial Intelligence: How EU regulation will impact the global AI market* (August 2022), Centre for the Governance of AI, at 15, https://uploads-ssl.webflow.com/614b70a71b9f71c9c240c7a7/630534b77182a3513398500f_Brussels_Effect_GovAI.pdf [hereinafter, Siegmann & Anderljung, *The Brussels Effect and AI*].

systems would be used in one of eight specified categories; and (2) the risk to health, safety, or fundamental rights is equal to or greater than the risk of harm from existing listed high-risk systems.[73] This system thus may not be able to quickly incorporate prohibitions of extreme high-risk AIs or to quickly add high-risk AIs that do not satisfy the two above conditions. Importantly, a risk-based global governance regime relating to AIs posing existential risks should ensure its classification of tiers of risks is amendable and does not establish needlessly sharp cliffs between tiers.

Notably, international treaties regulating dual-use substances have also taken a risk-based approach, such as the Vienna Convention on Psychotropic Substances (VCPS).[74] Just as some AIs offer both beneficial and harmful applications, some psychotropic substances have both beneficial and harmful applications.[75] Also, like a prior draft definition for AI in the EU AI Act,[76] the VCPS takes a risk-based approach to scope, defining a regulated "psychotropic substance" as "any substance, natural or synthetic, or any natural material in Schedule I, II, III, or IV."[77] With Schedule I being the most restrictive classification, regulatory requirements apply to substances based on their risk tiers; also, regulated substances are not determined by criteria but are rather specifically enumerated.[78] If needed, the VCPS permits the Commission on Narcotic Drugs of the Economic and Social Council of the United Nations to add a substance, following a determinative, scientific finding from the World Health Organization.[79]

Also, in another global regime for dual-use substances, the Montreal Protocol on Substances that Deplete the Ozone Layer (Montreal Protocol) reflects the international use of a largely risk-based approach to effectively regulate substances that belong to a similar grouping but have vastly different potential for harm.[80] For its scope, the Montreal Protocol defines "controlled substances" as those listed in its annexes, specifying it includes those "existing alone or in a mixture . . . [and] isomers of any such substance, except as specified in the relevant Annex, but excludes any controlled substance or mixture which is in a manufactured product other than a container used for the transportation or storage of that substance."[81] Thus, this definition reflects the risk-based, enumerated approach, including exemptions to reduce over-inclusiveness. For its enumerated substances, the Montreal Protocol offers listed categories and varying requirements in its annexes with different exemptions, as well as listed products

---

[73] EU AI Act, Nov. 2022 Draft, Title III, ch. 1, art. 7(1).
[74] **[Bluebook Cite]** [hereinafter the VCPS].
[75] Nexus of Hope, *The Benefits and Risks of Taking Psychiatric Medications*, https://nexusofhope.com/the-benefits-and-risks-of-taking-psychiatric-medications/#:~:text=They%20also%20affect%20your%20feelings,serotonin%20%E2%80%93%20your%20happy%20hormone.
[76] EU AI Act, April 2021 Draft art. (3)(1).
[77] VCPS, art. 1(e).
[78] See *id.* Schedules I–IV.
[79] See *id.*, art. 2.
[80] The Montreal Protocol on Substances that Deplete the Ozone Layer to the Vienna Convention for the Protection of the Ozone Layer, https://ozone.unep.org/sites/default/files/2019-12/The%20Ozone%20Treaties%20EN%20-%20WEB_final.pdf [hereinafter Montreal Protocol].
[81] Montreal Protocol, art. 1(4).

containing such substances.[82] Aside from amending the Montreal Protocol as detailed in Article 9 of the Convention for the Protection of the Ozone Layer, parties wishing to change the substances in the annexes must conduct an assessment before making any adjustments.[83]

Further, while it defines its scope based on a term's definition, the Cartagena Protocol on Biosafety also relies on a form of risk-based regulation in that it incorporates a risk assessment process in one of its core governance mechanisms.[84] To define its scope, the Cartagena Protocol defines the terms "living modified organism"[85] and states it will apply to the "transboundary movement, transit, handling and use of all living modified organisms that may have adverse effects on the conservation and sustainable use of biological diversity, taking also into account risks to human health."[86] In other words, the scope is determined based on the definition, which hinges on risks to conservation, to sustainable use of biological diversity, and to human health. Despite its reliance on this definition, the Cartagena Protocol does heavily rely on assessments of risk to inform decisions, providing state parties guidance on objectives, use of such assessments, general principles, methodology, and factors to consider in conducting risk assessments.[87]

To summarize, **Table 1** lists approaches treaties have adopted to define their regulatory scope.

Table 1: Regulatory scope approaches in international law

Regulatory scope approach	**Example**
Definition-based scope	IAEA Statute, art. II.
Definition-based scope incorporating risk assessment	Cartagena Protocol, art. 4; annex III.
Risk-based scope with tiers based on enumerated substances	Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES), art. 1(b).
Risk-based scope with tiers based on enumerated substances and necessary characteristics of such substances	Basel Convention on the Control of Transboundary Movements of Hazardous Wastes and Their Disposal, art. 1.

Thus, given the wide-spread acceptance of risk-based approaches to regulation in international treaties, a risk-based approach to global AI governance would not be unusual.

[82] Montreal Protocol, Annex A, B, C, D, E, & F.
[83] Montreal Protocol, art. 2(9), 2(10), and 6.
[84] Cartagena Protocol on Biosafety to the Convention on Biological Diversity [hereinafter the Cartagena Protocol].
[85] Cartagena Protocol art. 3(g) ("any living organism that possesses a novel combination of genetic material obtained through the use of modern biotechnology").
[86] Cartagena Protocol art. 4.
[87] Cartagena Protocol Annex III.

Framing risk tiers for AIs posing existential risks will likely require flexibility for updating such descriptions,[88] as well as a mixture of Schuett's described technical approaches,[89] applications, and capabilities to determine the scope of regulation.[90]

In line with this risk-based approach, the proposed governance entities and mechanisms in Parts II and III may be viewed as an inverse implementation of Robert Baldwin and Julia Black's explanation of a certain risk-based system–where a risk is framed based on the threat it poses to an objective.[91] Baldwin and Black acknowledge this method of "moving from a statement of statutory objectives to a set of key risks . . . is not a mechanical process . . . [and] involves a host of discretionary and value-laden decisions."[92] Instead of taking this approach, I assess well-known risks from AIs that may pose such existential risks *before* determining objectives that address those risks, using well-studied risks and frameworks to relatively minimize discretion in deciding objectives. From these objectives, I then explore mechanisms to accomplish those objectives, as well as legal entities for implementing those mechanisms.

II. Global Protective Safety Measures for High-Risk AI

This Part provides a brief introduction to a key source of existential risk from AI: the alignment problem. It then outlines possible methods of approaching future possible harms from unaligned, high-risk AI, including the concepts of risk, systemic risk, the precautionary principle, and uncertainty. With this toolkit of approaches to high-risk AI, this Part then explores initial possible protective safety measures that would mitigate risks caused in part by unaligned AI.

Further, this Part proposes ideal characteristics of legal models for global governance of high-risk AI, analyzing 11 treaties based on the presence or absence of those characteristics. Finally, based on existing legal models, this Part recommends possible components of a global safety system addressing high-risk AI.

A. The alignment problem

This section provides an overview of sources of existential risks that necessitate global safety measures for high-risk AI, focusing on the alignment problem. It is not exhaustive; rather, it seeks to explain the alignment problem and possible ways that it can be mitigated by global safety measures.[93]

---

[88] *See* n. 40 (describing the incredible growth in ML chips' speed and algorithmic efficiency for model training).
[89] RLHF and constitutional AI are currently dominant technical approaches in fine-tuning AIs that may become AGIs, but new manners of fine-tuning are likely to emerge. Targeting foundation models built with these methods may thus become underinclusive over time. See *supra* Part I.A.
[90] Relying too heavily on capabilities to frame risk would be a mistake–humans are currently not good judges of an AI's capabilities. See *infra* Part II.A (describing deception and sudden, emergent capabilities from AI).
[91] Baldwin & Black, *supra*, note 56, at 573–74.
[92] *Id.* at 582.
[93] For literature describing various sources of existential risk from AI, see *supra*, note 7.

A key source of existential risk from AI is the unsolved alignment problem. While described in different ways, the alignment problem may be framed as an intent problem–the difficulty of "building powerful AI systems that are aligned with their operators," meaning AIs that will try to do what their operators want them to do.[94] Given that AIs are trained to produce outputs based on optimization of an objective function, such as a reward function in RLHF, the outer alignment problem describes the difficulty of implementing an objective function that describes the desired behavior from the AI, while also not rewarding misbehavior.[95] Further, even if outer alignment is solved, the inner alignment problem would need to be addressed–AIs might develop subgoals that differ from those conveyed by the objective function.[96] For instance, during training, certain subgoals–such as gathering resources, information, or power–may enable an AI to consistently score highly on an objective function, possibly resulting in an AI with a goal of acquiring power.[97]

Also, inner alignment and outer alignment combined do not completely encompass the intent problem.[98] Even with a "'safe' objective function,"[99] the "'intention,' 'incentive,' or 'motive'" of an AI cannot be equated to the objective for which it is optimized.[100] Humans, for instance, are arguably optimized for the objective of reproductive fitness, yet few, if any, humans act with intent to optimize their reproductive fitness.[101] Along with problems arising from lack of intent alignment, AIs have demonstrated other problematic behavior, such as emergent functionality of new capabilities or goals that spontaneously emerge, as well as deception of human evaluators.[102] Given these unsolved and poorly understood risks from AI, the emergence of more powerful, capable AIs that mimic existing problematic behavior of less-capable AIs would pose extreme risks.[103]

Despite these obvious, well-known dangers, AI labs in the United States are openly competing to develop AGI and are raising increasing amounts of funding toward this goal.[104] If the needed capabilities to train foundation models spread outside of the United States and lead to an AGI race among states, these race dynamics may increasingly lead to a race to the bottom, as developers would be incentivized to win the race by shedding safety precautions. Absent widespread voluntary adoption of safety measures or a global agreement to abide by such measures, AI researcher Richard Ngo argues the culmination of this development is likely to

---

[94] Paul Christiano, *Clarifying "AI alignment*" (April 7, 2018), https://ai-alignment.com/clarifying-ai-alignment-cec47cd69dd6.
[95] Richard Ngo, *AGI Safety from First Principles* (Sep. 2020), at 18, https://drive.google.com/file/d/1uK7NhdSKprQKZnRjU58X7NLA1auXlWHt/view.
[96] *Id.* at 19. In other words, outer alignment is "the problem of correctly evaluating AI behavior; inner alignment is the problem of making the AI's goals match those evaluations." *Id.* at 21.
[97] *Id.* at 19.
[98] *Id.* at 21.
[99] *Id.* at 21
[100] *Supra*, note 50.
[101] *Id.*
[102] Hendrycks & Mazeika, *X-Risk Analysis, supra* note 7, at 5.
[103] *Id.* at 1.
[104] *See supra*, note 3.

result in humanity becoming a "second species"–one that cedes control of its future to unaligned AGIs.[105]

At a minimum, protective safety measures can help slow down this dangerous race toward unaligned AGIs or at least buy humanity more time to research and hopefully solve the alignment problem and other AI safety issues before an AGI is developed; at best, the safety measures can ensure safety by stopping certain dangerous developments altogether.[106]

B. Frameworks for approaching unaligned, high-risk AI

To better understand possible approaches to safety measures under risk-based governance, this section examines various methods of assessing unaligned, high-risk AI, including risk, systemic risks in the context of complex systems, the precautionary principle, and uncertainty frameworks. These frameworks are not mutually exclusive and offer a useful toolkit for global policymakers.

1. Risk

Risk frequently refers to a quantifiable, possible future harm. Margot Kaminski notes that risk is defined as the "possibility of loss or injury" and has roots in insurance, being "something to be measured, often mitigated, and taken into account."[107] She further comments that a formal risk analysis often involves calculations–with risk being calculated by multiplying the likelihood of an event and the measurable harm to be caused by such an event.[108] In the AI existential risk literature, risk has been framed as the following: Risk = (Hazard Severity and Prevalence) x (Exposure) x (Vulnerability).[109]

Given that risks often involve threats to be measured, the framing of the risk is crucial for subsequent risk-based regulatory regimes.[110] When approaching possible future harms through the risk framework, policymakers tend to be utilitarian; focused on quantifiable harms; focused on harms that result from muddled chains of causality and that are often externalities to those producing them; and focused on possible harms arising out of actions undertaken with the potential for a beneficial outcome.[111] In other words, these common traits of the risk framework

---

[105] Ngo, *AGI Safety from First Principles, supra*, note 51, at 1.
[106] *See Pause Giant AI Experiments: An Open Letter* (Mar. 22, 2023), Future of Life Institute, https://futureoflife.org/wp-content/uploads/2023/05/FLI_Pause-Giant-AI-Experiments_An-Open-Letter.pdf.
[107] Margot E. Kaminski, *Regulating the Risks of AI*, 103 B.U. L. Rev. __, 6 (2023) (forthcoming) [hereinafter Kaminski, *Regulating the Risks of AI*].
[108] *Id.* at 8.
[109] This framing defines a *hazard* as a "source of danger with the potential to harm," weighted with a hazard's probability of occurring; *exposure* as the "extent to which elements (*e.g.*, people, property, systems) are subjected or exposed to hazards"; and *vulnerability* as "susceptibility to the damaging effects of hazards." Hendrycks & Mazeika, *X-Risk Analysis*, *supra*, note 7, at 2.
[110] Baldwin & Black, *supra*, note 56, at 569.
[111] Kaminski, *Regulating the Risks of AI*, *supra* note 104, at 8–9.

may impose certain limitations, such as the difficulty of mitigating problems that are difficult to quantify through the risk framework.[112]

Further, risk-based frameworks follow a set process for the construction of an oversight regime. Typically, the regulating entity does the following: (1) set the level and types of risk it will tolerate; (2) conduct a risk assessment, including the likelihood of the harm occurring; (3) "evaluate the risk and rank the regulated entities [or activities] on their level of risk–high, medium or low;" and (4) allocate resources based on the tiers of risk.[113]

Following this process, the current structure of the EU AI Act reflects that its drafters: (1) determined tolerable levels of risk relative to threats to public interests and EU fundamental rights; (2) conducted risk assessments of the likelihood of harms from enumerated AIs or specific AI applications; (3) ranked those enumerated AIs and applications based on unacceptable, high-risk, limited risk, and minimal risk tiers; and (4) crafted regulations for each tier, such as conformity assessments for certain high-risk AIs.[114]

In the context of existential risks from unaligned AIs, applying this risk framework results in a few challenges. By definition, an existential risk involves a possible low-probability, high-impact event that may be difficult to accurately estimate.[115] For such rare or unprecedented events, quantifying such risks poses difficulties for knowledge, measurements, and mitigation.[116]

Even with such limitations, the risk framework is a useful base for constructing global AI regulatory safety measures, including to address the alignment problem. First, the parties to such an intergovernmental system ("the Parties") would need to determine the level and types of risk they are comfortable tolerating from unaligned, high-risk AI. In defining existential risks to be addressed, the Parties would also need to be careful to not needlessly cabin regulation within a specific sector or AI application.[117] The framing of the types of risk to be addressed is subjective and heavily influenced by policy preferences,[118] but this system could theoretically focus on risks to life and human health from unaligned AI, determining certain likelihoods of widespread catastrophe or death to be unacceptable or tolerable to a limited extent.

Second, during a risk assessment to determine the likelihood of existential harms to life or health from unaligned AIs, the Parties may struggle to determine the "specific details" of *how* such a catastrophe may occur, but the high probability of such an accident may be easier to

---

[112] *Id.* at 32–35 (noting harms that are not quantifiable or difficult to quantify without arbitrary policy choices).
[113] Michael Guihot, Anne F. Matthew, & Nicholas P. Suzor, *Nudging Robots: Innovative Solutions to Regulate Artificial Intelligence*, 20 Vanderbilt J. of Entertainment & Technology L. 385, 450 (2017) [hereinafter, Guihot et al., *Nudging Robots*].
[114] EU AI Act, Nov. 2022 Draft; Siegmann & Anderljung, *The Brussels Effect and AI*, *supra* note 70, at 15.
[115] Hendrycks & Mazeika, *X-Risk Analysis*, *supra*, note 7, at 1 (noting some single-digit estimates for the occurrence of AI-posed existential risks in the next century).
[116] Kaminski, *Regulating the Risks of AI*, *supra* note 104, at 34. Despite these difficulties, many are beginning to attempt such estimates. *See, e.g.*, Toby Ord, The Precipice 167 (2020) (estimating the likelihood of an existential catastrophe from unaligned AI within the next 100 years as 10%). **[something about forecasting]**.
[117] Alicia Solow-Niederman, *Administering Artificial Intelligence*, 93 S. Cal. L. Rev. 633, 670 (2020) ("cabining [AI] regulation too firmly within particular sectors splinters imperative conversations about cross-cutting values and norms for all domains.").
[118] *See* Part I.B.

estimate.[119] Currently, any highly capable AI is guaranteed to be unaligned to an extent.[120] Thus, assuming continuous increases in the quality and quantity of algorithms, ML chips, and compute dedicated to developing an AI, some of the prominent unknowns for this risk assessment are *if* and *when* a sufficiently capable, unaligned AGI can be or will be created.

Given sparse data on existential risks from AI, Schuett's factors for defining tiers of risk, especially technical approaches regarding the "making" of an AI and applications of such AIs, are useful proxies for estimating existential risks to life and health.[121] Assuming an AI will be unaligned, likelihood of risk would largely rely on estimates of an AI's projected capabilities.

Thus, risk assessments–and risk tiers built on these assessments–could largely rely on an inputs-based, technical approach to estimate the likelihood of risk from a future foundation model, as inputs like compute, ML chip quality, and training run time are good indicators of an AI's capabilities.[122] Initial input-based risk assessments would likely need to be calculated and planned by a multidisciplinary team of experts, resulting in risk tiers that have built-in adjustments accounting for technological progress or mechanisms to account for progression of ML chips, algorithms, and other inputs, along with a built-in margin of safety.

Third, the Parties would evaluate the risks and rank regulated entities and/or activities, creating tiers. Fourth, each of those tiers would ideally receive a tailored global treatment, including possible bans for unaligned AIs deemed too capable or mandatory safety practices.

2. Systemic risk

Risk can also be characterized from the perspective of systemic or structural factors that increase the probability of a harmful future outcome.[123] Instead of framing risk as a singular entity, Baldwin and Black contend risk is more usefully viewed "as a cluster of different causes and effects that is assembled for a given purpose according to a principle of framing or selection."[124] For instance, in studying existential risks, many no longer characterize such risks "as singular events and . . . [prefer] descriptions of risks as the result of the complex interactions between multiple, more mundane vulnerabilities in our social and political systems."[125] Hendrycks and Mazeika note that "modern systems are replete with nonlinear causality, including feedback loops, multiple causes, circular causation, self-reinforcing processes, butterfly effects, microscale-macroscale dynamics, [and] emergent properties . . . [making it best] to ask how various factors contributed to a failure."[126]

---

[119] *See* Alicia Solow-Niederman, *supra* note 115, at 663.
[120] *See* discussions *supra* Part I.A and Part II.A.
[121] *See supra* footnote 65.
[122] See Part I.A.
[123] *See* Guihot et al., *Nudging Robots*, *supra* footnote 111, at 416 ("Systemic risk is the embedded risk 'to human health and the environment . . . in a larger context of social, financial and economic risks and opportunities.'").
[124] Baldwin & Black, *supra*, note 56, at 570.
[125] Benjamin S. Bucknall & Shiri Dori-Hacohen, *Current and Near-Term AI as a Potential Existential Risk Factor*, at 3, https://arxiv.org/pdf/2209.10604.pdf.
[126] Hendrycks & Mazeika, *X-Risk Analysis*, *supra*, note 7, at 3–4.

Drawing on the above principles of systemic risk, systemic risks from unaligned, high-risk AI may be characterized at the level of the AI global ecosystem or at the level of the AI model itself. First, in assessing the global ecosystem for the development of powerful foundation models, regulators seeking to address risk may need to consider risk factors, such as state-state competition, state-corporation power relationships, AI interactions with nuclear threats, and more.[127] Risk factors from the pursuit of an AGI–which will be unaligned if current conditions persist–are complex to address individually, but mechanisms that would increase the costs of developing AGI unsafely for all participants may be useful.

Second, in assessing systemic risks at the level of a foundation model, others have identified prominent risk factors as arising from (i) an AI's internal characteristics, such as interpretability; (ii) the external environment's effect on an AI, such as access to the internet; and (iii) the AI's effect on the external environment, such as an AI influencing a nuclear weapon launch.[128] For a focus on the second category of an environment's effect on an AI, the complex system that is the AI could make it difficult to "predict how a particular intervention will unfold . . . [or] to specify a cause-and-effect relationship."[129] For instance, despite being a process meant to increase the safety of an AI, RLHF may initially result in negative consequences or outputs from a given model.[130]

Given a systemic view of existential risk from unaligned AIs, safety measures would likely help mitigate global risk factors. However, these measures themselves would need to be mindful of the various risk factors that may result from the intervention itself.

3. The precautionary principle

For situations involving difficulty quantifying risks or arriving at probabilities, the precautionary principle may also be invoked by global regulators of high-risk, unaligned AI. With a variety of manifestations, the precautionary principle can broadly be described as the idea that "preventative or remedial measures can, should, or must be taken when there is scientific uncertainty that an unacceptable hazard may occur."[131] Generally, versions of the precautionary principle may be interpreted to "(a) not justify regulatory inaction (a minimalist approach); (b) justify regulatory action even if cause and effect have not been proven (a median approach); or even (c) necessitate regulation until it is clear there is no danger of serious harm (a maximalist approach)."[132] Naturally, to be applicable, one must specify a version of the principle.

---

[127] *See* Bucknall & Dori-Hacohen, *supra* footnote 123.
[128] Schuett, *supra* note 49, at 18 (citing José Hernández-Orallo et al., *Surveying Safety-relevant AI Characteristics*, Proceedings of the AAAI workshop on Artificial Intelligence Safety (2019)).
[129] Alicia Solow-Niederman, *supra* note 115, at 673–74.
[130] See Charbel-Raphaël, *Compendium of problems with RLHF*, *supra* footnote 52.
[131] Grant Wilson, *Minimizing Global Catastrophic and Existential Risks from Emerging Technologies through International Law*, at 33, https://gcrinstitute.org/papers/006_international-law.pdf.
[132] Kaminski, *Regulating the Risks of AI*, *supra* note 104, at 23.

At its weakest, the precautionary principle states that "scientific uncertainty is an inappropriate reason *not* to take precautionary actions."[133] For example, the Cartagena Protocol cites Principle 15 of the Rio Declaration on Environment and Development, which states "Where there are threats of serious or irreversible damage, lack of full scientific certainty shall not be used as a reason for postponing cost-effective measures to prevent environmental degradation."[134] For high-risk, unaligned AI, it would be relatively uncontroversial to apply this weak version; lack of certainty of whether unaligned AGI is possible should not be a reason for postponing cost-effective safety measures. This conclusion, however, offers little guidance for *what* action to take.

Further, given the potentially extinction-level or catastrophic risks posed by unaligned AI, applying an aggressive, maximalist form of the precautionary principle to global regulation may be desirable in certain instances. This version can be described as requiring regulation when "there is a possible risk to health, safety, or the environment, even if the supporting evidence remains speculative and the economic costs of regulation are high."[135] To avoid application of this principle to all risks, this principle requires a "certain threshold of scientific plausibility" that harms from such risks may occur.[136]

For unaligned, highly capable AI, the conditions for the application of this strong form of the precautionary principle are applicable. First, a highly capable unaligned AI poses massive risks to the health, safety, and environment. Second, supporting evidence overwhelmingly shows foundation models are unaligned to an extent; key uncertainty only remains as to whether or not AGI is possible, as well as the extent of the harm that would result from the arrival of an unaligned AGI. Finally, economic costs of regulating the powerful AI labs developing foundation models are likely to be high.

In an even stronger framing of the precautionary principle that is useful for unaligned AI posing an extinction-level threat, Cass Sunstein has suggested an Irreversible Harm Precautionary Principle. In short, it states "[w]hen regulators lack information about the likelihood and magnitude of a risk, it makes sense to spend extra resources to buy an 'option' to protect against irreversible harm until future knowledge emerges."[137] The cost of this "option" is that of "delaying the decision until better information is available."[138] Given the above-mentioned lack of understanding about the possibility of an unaligned AGI, humanity could buy an option to protect against irreversible harm until the alignment problem is solved, preventing the development of AGI. This option's value would be that of tolerating existing suffering and harms that could possibly be alleviated or solved by the creation of a powerful AI. However, it seems extremely unlikely that the United States and other necessary countries would completely ban development of unaligned foundation models with the *potential* to become AGI.

---

[133] Grant Wilson, *Minimizing Global Catastrophic and Existential Risks from Emerging Technologies through International Law*, at 34, https://gcrinstitute.org/papers/006_international-law.pdf.
[134] Cartagena Protocol (citing the Rio Declaration on Environment and Development (1992)).
[135] Cass R. Sunstein, *Irreversible and Catastrophic*, 91 Cornell L. Rev. 841, 850 (2006).
[136] *Id.*
[137] *Id.* at 845.
[138] *Id.*

4. Uncertainty, risk, and the precautionary principle

In situations where probabilities may not be assigned to a possible future harm, an approach based on uncertainty may also prove useful. *Uncertainty* refers to situations where no probabilities to possible outcomes may be assigned.[139] Even if one argues that true uncertainty is very rare or non-existent, "bounded uncertainty" is possible, meaning a situation where probabilities within bands of probability cannot be assigned.[140] Thus, even under an approach of uncertainty, policymakers may need to engage with the language of probability and risk.

In an approach for possibly catastrophic situations with uncertainty, maximin advocates for acting to eliminate the worst-case scenario or outcome. When possible outcomes are uncertain and different outcomes have the same best consequences, maximin advises to compare the worst consequences for different courses of action, choosing the course of action that has the relatively less worse consequence.[141] Comparing the best consequences of choosing to develop or not develop AGI is difficult, but assuming this first condition is present, the worst consequence of developing AGI is human extinction; the worst consequence of banning AGI development is likely a better outcome than extinction.

However, choosing to apply maximin is not that simple. Like when invoking the precautionary principle, applying maximin requires a minimum threshold of plausibility of the worst case scenario occurring–again, requiring the language of probability and risk.[142]

In a form of a cost-analysis approach common under a risk framework, Sunstein encourages policymakers to ask: "(a) How bad is the worst-case scenario, compared to other bad outcomes? (b) What, exactly, is lost by choosing maximin?"[143] Sunstein further argues that maximin is best when the gap between worst-case scenarios is very large and its application does not result in very high losses.[144] Again, if the calculated plausibility of human extinction from unaligned AI is sufficient, applying maximin meets the first criterion, although assigning probabilities of occurrence to each scenario being compared would also be necessary.[145] The second criterion is harder to calculate, but it is unlikely that applying maximin to unaligned AI will result in high losses.

Even if a complete ban on powerful AIs was implemented, it would need to be accompanied by other actions. While knowledge of AIs is likely to increase over time and enable

---

[139] *Id.* at 848.
[140] Cass R. Sunstein, *The Catastrophic Harm Precautionary Principle*, Issues in Legal Scholarship 20 (2007).
[141] Jon Elster, Explaining Technical Change: A Case Study in the Philosophy of Science 203 (1983).
[142] *See* footnote 134. A possible application of maximin would be to ban the creation of specified foundation models based on certain input ceilings.
[143] Sunstein, *Irreversible and Catastrophic*, at 889. In essence, Sunstein combines maximin and the precautionary principle to suggest his Catastrophic Harm Precautionary Principle, drawing attention to the need to assess losses imposed from applying any given regulation, along with the risks of inaction. Sunstein, *The Catastrophic Harm Precautionary Principle*, at 24.
[144] Sunstein, *Irreversible and Catastrophic*, at 892.
[145] Sunstein, *The Catastrophic Harm Precautionary Principle*, at 24.

better probabilities calculations for risks,[146] it is not clear scientists working at the forefront of new technologies are good predictors of when prominent capabilities of such technologies will manifest.[147] Further, an important component of addressing the alignment problem is ensuring that safety research results in a higher safety-capabilities ratio, meaning that research does not move AIs closer to becoming AGIs ("capabilities) more than it results in understanding of and progression toward overall safety.[148]

Even when framing maximin and the Catastrophic Harm Precautionary Principle in the language of risk and probabilities, arguments for not applying maximin do not apply to high-risk, unaligned AIs. Maximin should not be applied when the "worst case is highly improbable and when the alternative option is both much better and much more likely."[149] The worst case scenario from unaligned AI is estimated as about 10% in the next century by at least one researcher, so it is reasonable to state it is not *highly* improbable.[150] Further, most alternatives to extinction or mass catastrophe are clearly not better outcomes.

Finally, like the precautionary principle, Sunstein's Catastrophic Harm Precautionary Principle may be applied with varying degrees of force to unaligned AI. For instance, the weakest version of the Catastrophic Harm Precautionary Principle would have regulators consider the expected value of catastrophic risks and choose cost-effective measures to reduce those risks and maximize net benefits, "even when it is highly unlikely that those risks will come to fruition."[151] A second, stronger version of this principle would have regulators also consider the "social amplification of risk" from a catastrophe, as the initial expected value of the harm would likely not capture the full secondary losses from the deaths of millions.[152] Finally, an even more aggressive version of this principle would have regulators consider the expected value of catastrophic losses, social amplification risks, and the need for a margin of safety in imposing regulatory decisions.[153] Together, these principles provide policymakers with various options for approaching global regulation of unaligned AI.

C. Protective safety measures for unaligned, high-risk AI

[146] *Id.* at 27.
[147] *See* Ashutosh Jogalekar, *Leo Szilárd, a traffic light and a slice of nuclear history*, Scientific American, Feb. 12, 2013, https://blogs.scientificamerican.com/the-curious-wavefunction/leo-szilard-a-traffic-light-and-a-slice-of-nuclear-history/ (describing how prominent physicist Ernest Rutherford dismissed the idea of the release of energy from atoms as "moonshine" a mere six years before the discovery of nuclear fission).
[148] Hendrycks & Mazeika, *X-Risk Analysis*, *supra*, note 7, at 7–8.
[149] Sunstein, *Irreversible and Catastrophic*, at 879.
[150] See footnote 114.
[151] Sunstein, *The Catastrophic Harm Precautionary Principle*, at 3.
[152] *Id.* at 6.
[153] *Id.* at 8.

Building on the previous exploration of frameworks for assessing threats from unaligned, high-risk AI, this section explores possible initial protective safety measures–referred to as mechanisms.[154]

1. Substantive mechanisms

First, through a process known as red teaming, AI labs and other developers could be required to hire an external team to identify hazards in their high-risk AIs. For high-risk AI systems, red teaming would involve hiring external red teams to identify "hazards in . . . AI systems to inform deployment decisions, [such as exploring] dangerous behaviors or vulnerabilities in monitoring systems intended to prevent disallowed use."[155] Red teaming could also be used to "provide indirect evidence that an AI system might be unsafe."[156] In a survey of leading experts from AGI labs, academia, and civil society, 98% somewhat or strongly agreed that "AGI labs should commission external red teams before deploying powerful models."[157]

However, given the possible variety in backgrounds, skills, and resources of potential red teams, global minimal standards could help facilitate a floor of safety and quality, while also promoting rapid adoption of methods to overcome difficulties of effective red teaming. Possible requirements for red teams could include the need for (i) sufficient experience in interacting with state-of-the-art AI models; (ii) mandating a sufficient amount of access to the AI models being red teamed; (iii) ensuring adequate resources for red teams; (iv) mandating communication of certain results with regulators as needed; and (v) including red teams in the process of post-deployment model updates.[158] Further, given that red teaming may be increasingly difficult for certain RLHF models as they scale in size,[159] cutting-edge methods may be necessary to ensure effectiveness, such as red teaming possibly unaligned AIs with aligned or less capable AIs.[160]

Currently, ARC Evals, a prominent red teaming organization, provides a preliminary model of possible methods for addressing threats posed by high-risk AIs. ARC Evals partners with prominent AI labs Anthropic and OpenAI, and the organization explicitly mentions its concerns about global catastrophe from misaligned AI on its home page.[161] In particular, ARC Evals is concerned with misaligned AIs' potential to autonomously replicate, concentrating its current red teaming efforts on evaluating existing models' autonomous replication abilities.[162] In

---

[154] [For a more complete list of technical mechanisms, see [link list].]

[155] Hendrycks et al., *An Overview of Catastrophic AI Risks*, at 32.

[156] *Id.*

[157] Jonas Schuett et al., *Towards best practices in AGI safety and governance: A survey of expert opinion*, at 18, https://arxiv.org/pdf/2305.07153.pdf.

[158] Anderjlung et al., *Frontier AI Regulation*, *supra* note __, at 26.

[159] Deep Ganguli et al., *Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned*, Anthropic, https://arxiv.org/pdf/2209.07858.pdf.

[160] Ethan Perez et al., *Red Teaming Language Models with Language Models*, Google DeepMind, https://arxiv.org/pdf/2202.03286.pdf.

[161] Arc Evals, https://evals.alignment.org/.

[162] *See id.*

a recent evaluation of models based on leading LLMs from Anthropic (Claude) and OpenAI (GPT-4), ARC Evals found it unlikely that such models were capable of creating dangerous autonomous agents.[163] This evaluation's report provides a preview of possible substantive components of red teaming, which may include assessing model capabilities based on a series of increasingly difficult tasks ranging from searching a filesystem for a password to creating a language model agent.[164]

Second, at the systems level of the global AI community, a significant substantive mechanism would be the implementation of safety education and practices that promote a global culture of responsibility and safety, particularly among regulators and leading AI labs. Global ethical standards and statements are already in existence, as demonstrated by the work of the IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems, as well as the Asilomar AI Principles of the Future of Life Institute.[165]

However, these guiding standards likely need to be reinforced by more stringent, binding standards applicable to high-risk AI development and deployment.[166] Researchers working with high-risk AI have also recognized the need for augmenting such a safety culture, even suggesting the establishment of a reporting structure or hotline to regulators to mitigate risks from highly capable models.[167] In the global context, such a reporting structure could be uniformly mandated by all countries where high-risk AI models are being developed, requiring certain activities to be reported to a respective national regulatory authority.

Third, given certain models' potential to demonstrate or be modified to produce new risks after deployment, uniform post-deployment obligations for providers of high-risk models could be standardized globally. Depending on how a model is released or accessible to users, new dangerous capabilities can be further developed or unlocked through (i) fine-tuning; (ii) chain-of-thought prompting, which involves telling a model to think through problems step by step to improve problem-solving capabilities; (iii) enabling LLMs to use external tools; (iv) automated prompt engineering, which involves "[u]sing LLMs to generate and search over novel prompts that can . . . elicit better performance on a task; and (v) foundation model programs, which integrate foundation models into complex programs.[168] Substantive post-deployment

---

[163] Beth Barnes, *ARC Evals new report: Evaluating Language-Model Agents on Realistic Autonomous Tasks* (Aug. 1, 2023), https://www.lesswrong.com/posts/EPLk8QxETC5FEhoxK/arc-evals-new-report-evaluating-language-model-agents-on.

[164] *Id.*

[165] Miles Brundage et al., *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*, Future of Humanity Institute, at 56, https://arxiv.org/pdf/1802.07228.pdf.

[166] *See* Matthew Hutson, *Who Should Stop Unethical A.I.?*, The New Yorker (Feb 15, 2021), https://www.newyorker.com/tech/annals-of-technology/who-should-stop-unethical-ai (noting the lack of ethical standards in computer science research and development); Hendrycks & Mazeika, *X-Risk Analysis for AI Research*, at 3–4, 9 (noting systemic risk factors relating to safety, such as safety culture, safety team resources, incentive structures, and more, as well as recommending empirical measurements of safety goals).

[167] Fabio Urbina et al., *Dual use of artificial-intelligence-powered drug discovery*, 4 Nature Machine Intelligence 191 (2022).

[168] Anderjlung et al., *Frontier AI Regulation*, *supra* note __, at 12.

obligations could involve regular risk assessments, provision of incident reporting mechanisms, and mechanisms for quickly withdrawing a deployed model.[169]

2. Procedural mechanisms

Along with substantive mechanisms, procedural mechanisms can help promote safety; for instance, a crucial *ex ante* mechanism to help control high-risk AI development or deployment could be the use of a globally mandated licensing process for high-risk models.[170] In the U.S. administrative law context, licensing involves an expert agency that "sets standards for and assesses the safety of a technology or practice" before its release to the public.[171] This assessment may be relative to whether the technology, practice, or an entity meets a performance standard, meaning this procedural mechanism incorporates a substantive threshold for safety.[172] For example, in both the European Union and the United States, financial institutions may require a license to operate and can have that license revoked.[173]

For global oversight of high-risk AI, the licensing process could require countries to license the deployment and development of high-risk AIs.[174] Deployment-based licensing would closely parallel licensing of pharmaceutical drugs, for example, granting market access only if a deployer could demonstrate compliance with specific safety standards.[175] Additionally, development-based licensing could provide an additional layer of safety, requiring licensing for certain activities or stages in the training process of a high-risk foundation model.[176]

Further, given the possibility of catastrophic harm from certain models, it may be desirable for global licensing standards to place the burden of proof on AI developers to demonstrate their models meet necessary safety standards.[177] Another desirable trait of a global licensing regime would be to establish conditional licensing, mandating ongoing compliance with specified safety guard rails.[178] However, in establishing a licensing regime, global regulators would need to be cautious about entrenching centralized power with regulated companies and AI labs.[179]

Also, in another procedural mechanism that is arguably a variation of a licensing system, a conformity assessment performed by a third party or regulated entity may be required before a

---

[169] *Id.* at 28.
[170] *See* Ho, *International Institutions*, at 6.
[171] Kaminski, *Regulating the Risks of AI*, at 28; Andrew Tutt, *An FDA for Algorithms*, 69 Admin. L. Rev. 83, 111 (2017) (recommending a system where an agency must approve an algorithm before its public release).
[172] Andrew Tutt, *An FDA for Algorithms*, 69 Admin. L. Rev. 83, 111 (2017)
[173] Anderjlung et al., *Frontier AI Regulation*, *supra* note __, at 19.
[174] *Id.* at 20.
[175] *Id.*
[176] *Id.*
[177] *See* Gianclaudio Malgieri & Frank Pasquale, *From Transparency to Justification: Toward* Ex Ante *Accountability for AI*, Brooklyn Law School, Legal Studies Paper No. 712 (2022).
[178] Kaminski, *Regulating the Risks of AI*, at 83.
[179] Anderjlung et al., *Frontier AI Regulation*, *supra* note __, at 31.

model is released, as well as after its release.[180] For example, the EU AI Act requires a successful conformity assessment from third-party private entities for its designated high-risk AI models, which upon successful completion may be designated as sufficiently safe through an "EU technical documentation assessment certificate."[181] The EU AI Act conformity assessment mandates compliance with requirements for a risk management system, record-keeping, human oversight, and post-market monitoring, among others.[182] While a conformity assessment conducted by a non-governmental entity may be a useful complement to government licensing, these private entities are not accountable to the public in the same manner many democratically elected governments are.

Finally, a procedural mechanism aimed toward safety could be the requirement for a pre-publication risk assessment before the release of potentially dangerous information in a working paper or scholarly article tied to a high-risk model. This pre-publication risk assessment could mandate analysis of "particular risks . . . of a particular capability if it became widely available, and deciding on that basis whether, and to what extent, to publish it."[183] For instance, the publication of a highly capable model's parameters would enable its fine-tuning for possibly nefarious purposes, and a publication review could highlight such risks, among others.[184]

D. Ideal characteristics of legal models for global AI safety measures

To explore possible legal models for an international system implementing protective safety mechanisms, this section seeks to identify relevant, desirable characteristics. It further preliminarily examines to what extent existing international treaties or intergovernmental bodies possess these characteristics.

Dual-use technology or substances. A first characteristic that could be used to identify ideal legal frameworks for global high-risk AI governance is a focus on frameworks useful for governance of dual-use technologies.[185] Based on a technology's inherent characteristics, its dual-use nature can be defined based on its possible application for both military and civilian purposes, such as in commercial settings.[186] On the other hand, based on a technology's externalities, its dual-use nature can be framed as its potential to be used for harmful purposes or

---

[180] Kaminski, *Regulating the Risks of AI*, at 52 (describing the conformity process as "licensing lite" conducted by third parties or the regulated entities themselves).
[181] *Id.*
[182] Siegmann & Anderljung, *The Brussels Effect and AI*, *supra* note 70, at 15.
[183] Miles Brundage et al., *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*, Future of Humanity Institute, at 86, https://arxiv.org/pdf/1802.07228.pdf.
[184] *See* Part I.A; Hendrycks et al., *An Overview of Catastrophic AI Risks*, *supra*, note 63, at 32.
[185] "[T]he concept of dual-use technology is a useful organizing principle for examining governance efforts across different technology areas." Elisa D. Harris et al., Introduction: Governance of Dual-Use Technologies: Theory and Practice, American Academy of Arts & Sciences, https://www.amacad.org/publication/governance-dual-use-technologies-theory-and-practice/section/3 [hereinafter Harris et al., *Governance of Dual-Use Technologies*].
[186] *Id.* (citing Council Regulation (EC) No. 428/2009 of May 5, 2009, Official Journal of the European Union (May 29, 2009): L134/3, https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:134:0001:0269:en:PDF and Dual Use Exports, C.F.R., Title 15, § 730.3 (2000)).

human betterment.[187] For this characteristic, I use the latter definition, considering a technology's potential military purpose as harmful and civilian purposes as largely contributing to human betterment. Regardless, under either of these definitions, AI is a dual-use technology.

Risk-based governance scope. Second, as previously defined in Parts I.B. and II.B, a risk-based scope for governance of substances and technologies is an effective approach for unaligned AI. To be clear, by a "risk-based scope," I refer to a regulatory regime that regulates specified similar substances or groups of technologies based on tiers of their respective risks.[188] [However, I do note regulatory regimes that base their scope on a term's definition, while also incorporating risk assessments.]

Provision of safety measures. Third, given an objective of providing, updating, and enforcing safety measures, another ideal characteristic is that the organization contributes measures promoting safety.[189] In evaluating whether a comparable treaty provides "safety measures," I assess whether or not the treaty incorporates substantive requirements or procedural processes that mitigate against the specific risks the treaty's objective seeks to address.

Factors from AI experts (Anderjlung Factors). Further, another useful characteristic of an AI governance regime is possessing certain features noted by Markus Anderjlung and other AI experts, including (a) including expert-driven, multi-stakeholder processes, (b) being capable of rapid iteration, and (c) relying on technically-informed processes.[190] To the extent that other governance regimes share these characteristics, their structures may be useful in designing a global governance regime for high-risk AI.

Effects consistent with a treaty's intent. Further, another pertinent characteristic for any international legal regime is that it produce its intended effects. In a prominent meta-analysis of 82 studies on international treaties, Hoffman et al. found evidence that many treaties mostly failed to produce their intended effects, except for (a) trade and finance treaties and (b) treaties incorporating enforcement mechanisms.[191] To supplement these quantitative analyses, Hoffman et al. provide qualitative data of other studies evaluating treaties for whether or not they produced their intended effects.[192] To determine if a given treaty possesses the "intended effects" characteristic, I use both the (a) meta-analysis determination for a specific treaty and (b) qualitative studies cited in the Hoffman appendix.

Despite the usefulness of the Hoffman study, its limitations are also important to consider, suggesting the intended effects characteristic should not alone be dispositive or even overly weighted in importance. For instance, the meta-analysis is limited by the quantity and

---

[187] Harris et al., *Introduction: Governance of Dual-Use Technologies*.
[188] *See* Part I.B. (describing Vienna Convention on Psychotropic Substances and the Protocol on Substances that Deplete the Ozone Layer).
[189] Harris et al., *Introduction: Governance of Dual-Use Technologies* (noting that dual-use governance regimes often have the objective of "promoting the safe and secure handling and use of materials, equipment, or information associated with dual-use technologies"); *see also* Anderjlung et al., *Frontier AI Regulation*, at 23–29.
[190] Anderjlung et al., *Frontier AI Regulation*, at 3, 16, 21.
[191] Steven J. Hoffman et al., *International treaties have mostly failed to produce their intended effects* (Aug. 1, 2022), https://www.pnas.org/doi/full/10.1073/pnas.2122854119#t02.
[192] Steven J. Hoffman et al., *Supplementary Information for International Treaties Have Mostly Failed to Produce their Intended Effects*, Table S9.

quality of its included primary studies, and it is possible Hoffman's study missed relevant primary studies.[193] Further, despite its finding of the economic effectiveness of trade and finance treaties, it is possible such outcomes are to be expected, as such treaties may be "more easily measured, produce more easily quantified effects, or are more consistently studied using high-quality quantitative methods" than other treaties.[194] Finally, Hoffman et al. encourage caution in interpreting causal interpretations of intended and unintended treaty impacts.[195]

Enforcement mechanisms. Finally, given the Hoffman analysis's finding of the usefulness of enforcement mechanisms, an important characteristic of a global regulatory regime for high-risk AI is the presence or absence of such mechanisms. An enforcement mechanism facilitates "the possibility of a specific sanction or consequence delivered by a court, committee, secretariat, or other legal authority, even if that authority was not created by the treaty."[196] For example, specific sanctions or consequences include "financial sanctions on countries or expelling countries from treaty bodies and trade blocs."[197] I include this factor as a stand-alone characteristic to account for treaties that may not have been included in the Hoffman meta-analysis determination or the qualitative data of other studies in its appendix.

Given these characteristics of an ideal global regulatory regime for high-risk AI, I provide below in **Table 2** a preliminary analysis of which pertinent international treaties possess the specified characteristics.

Table 2: International Treaties and Ideal Characteristics of Global Governance of High-Risk AI

Applicable Treaty or Institution	Dual-use nature	Risk-based governance scope	Provision of safety measures	Anderjlung Factors	Intended Effects (Hoffman Meta-Analysis (H) & Qualitative List (QL))	Enforcement mechanism
Single Convention on Narcotic Drugs (SCND), 1961	Yes	Yes	Yes	(a) Yes (b) Yes (c) Yes	H: N/A QL: N/A	Yes, art. 14 provides for Board oversight and art. 48(2) for ICJ referral.

[193] See *supra*, note 192.
[194] *Id.*
[195] *Id.*
[196] *Id.*
[197] *Id.*

Vienna Convention on Psychotropic Substances (VCPS), 1971	Yes	Yes, art. 1(e).	Yes	(a): Yes (b): No (c): Yes	H: N/A QL: N/A	Yes. International Court of Justice (ICJ). art. 31
Convention Against Illicit Trafficking of Narcotic Drugs and Psychotropic Substances (CAIT), 1988	Yes	Yes. art. 1(n)	Yes	(a) Yes (b) No (c) No	H: N/A QL: N/A	Yes, art. 32.
Montreal Protocol on Substances that Deplete the Ozone Layer (Montreal Protocol)	Yes	Yes, art. 1(4).	Yes, such as art. 7 (data reporting) & art. 4(b) (licensing)	(a): Yes (b): No (c): Yes	H: N/A QL: Yes (ID 449)	Yes (pg. 818), up to suspension by Parties.
Cartagena Protocol on Biosafety to the Convention on Biological Diversity (Cartagena Protocol)	Yes	No, definition-based scope, but includes risk assessment process	Yes, art. 18.	(a): Yes (b): No (c): Yes	H: N/A QL: N/A	Yes, under art. 27 of the Convention on Biological Diversity.
Biological Weapons Convention (BWC)	Yes, but art. I excludes those used for peaceful purposes	No	Yes	(a): No (b): No (c): No	H: No QL: No	No, but referral to the UN Security Council (art. VI)
Treaty on the Non-Proliferation of Nuclear Weapons (NPT)	Yes	No	Yes, art. III(2).	(a): No (b): No (c): No	H: Yes QL: Not effective (not opposite effect)	No
Convention on	Yes	No	Yes	(a): Yes	H: N/A	No

Nuclear Safety[198]				(b): No (c): Yes	QL: N/A	
Convention on International Trade in Endangered Species (CITES)	No	Yes	Yes, art. VIII provides measures to prevent specimens from being traded	(a): Yes (b): No (c): Yes	H: Yes QL: Yes (ID 364); No (ID 377)	Yes, Permanent Court of Arbitration.
Basel Convention on the Control of Transboundary Movements of Hazardous Wastes and their Disposal	No	Yes, risk-based tiers based on categories, characterist -ics and domestic legislation. art. 1	Yes. art. 4	(a): Yes (art. 10) (b): No, but art. 17 provides for amendment s by ¾ majority (c): Yes	H: N/A QL: No (ID 578, ID 101)	Yes, but Parties may refuse to submit a dispute to the ICJ or to arbitration. art. 20.
Bern Convention on the Conservation of European Wildlife and Natural Habitats (Bern Convention)	No	Yes	Yes. *See, e.g.*, art. 5.	(a): Yes (b): No (c): Yes	H: N/A QL: N/A	Yes, art. 18 provides for arbitration.

E. Building a Global Safety Regime for High-Risk AI

From a comprehensive review of the above treaties and their global organs, I next focus on the components that may provide foundational blocks for the construction of an international system dedicated to addressing safety concerns from unaligned, high-risk AI, including a primary body: a Global Organization for High-Risk Artificial Intelligence (GOHAI).[199] In

[198] The international nuclear safety regime includes many treaties and agreements, including the NPT, the Convention on Nuclear Safety, the International Atomic Energy Agency (IAEA) Statute, and bilateral individual country agreements with the IAEA, among others. James M. Acton, *Governance of Dual-Use Technologies: Theory and Practice, Ch. 1: On the Regulation of Dual-Use Nuclear Technology*, https://www.amacad.org/publication/governance-dual-use-technologies-theory-and-practice/section/4.

[199] GOHAI may also be a division of an international body generally addressing AI, not just solely high-risk AI. For instance, along with its policymaking organs and staff, the IAEA has its Department of Management, Department of Technical Cooperation, Department of Nuclear Energy, Department of Nuclear Safety and Security, Department of

drawing on existing regulatory regimes to suggest components of GOHAI and a global regulatory system, I also aim to mitigate against what Kaminski terms as "classification conflicts," meaning issues that arise from relying on a "transplant of risk regulation methods and tools from other legal fields, often with widely varying institutions of origin."[200]

Possible components of GOHAI and affiliate organizations may include: (1) enforcement mechanisms; (2) a Secretariat and Depositary; (3) a financial mechanism to provide funding; (4) individual Parties' specialized bodies of experts; (5) a global impartial body of experts; (6) a global body of experts with the mandate to advocate for the execution of the treaty; (7) representative entities that oversee the implementation of a treaty, including updating risk tiers for high-risk AI; (8) a process for updating high-risk models' risk tier classifications; and (9) a representative body of all parties, including possible subsidiary bodies.

1. Enforcement mechanisms

Given the importance of treaties' enforcement mechanisms to ensure their effectiveness,[201] GOHAI should include enforcement mechanisms that will ensure compliance or at least substantially increase the cost of non-compliance for state Parties. An "enforcement mechanism" refers to a "specific sanction or consequence delivered by a court, committee, secretariat, or other legal authority, even if that authority was not created by the treaty."[202] For example, a treaty's enforcement mechanism may enable referral of disputes to an established judicial body, such as the International Court of Justice (ICJ)[203] or the Permanent Court of Arbitration.[204] Similarly, other treaties permit the establishment of an ad-hoc arbitration panel or conciliation committee for the resolution of disputes.[205]

Along with referring a dispute to a judicial entity, treaties may also permit parties to suspend the rights and privileges of other parties in certain circumstances.[206] Further, while not an enforcement mechanism, a mechanism that may eventually result in a particular sanction or consequence is the ability of parties to file complaints with the United Nations Security Council.[207]

Nuclear Sciences and Applications, and Department of Safeguards. IAEA Organizational Structure, https://www.iaea.org/about/organizational-structure.

[200] Kaminski, *Regulating the Risks of AI*, at 73 (citing Vanessa Casado Pérez & Yael R. Lifshitz, *Natural Transplants*, 97 N.Y.U. L. Rev. 933 (2022)).

[201] Steven J. Hoffman et al., *International treaties have mostly failed to produce their intended effects* (Aug. 1, 2022), https://www.pnas.org/doi/full/10.1073/pnas.2122854119#t02.

[202] *Id.*

[203] **[Bluebook cite fix]** *See, e.g.*, SCND art. 48(2); VCPS art. 31(2); CAIT art. 32(2); Vienna Convention for the Protection of the Ozone Layer (VCPOL) art. 11(3)(b); Convention on Biological Diversity (CBD) art. 27(3)(b); Basel Convention on the Control of Transboundary Movements of Hazardous Wastes and their Disposal (Basel Convention) art. 20; the IAEA Statute art. XVII.

[204] *See, e.g.*, CITES art. XVIII.

[205] *See, e.g.*, SCND art. 48(1); VCPS art. 31(1); CAIT art. 32(1); VCPOL art. 11(3)(a); CBD art. 27(3)(a); Basel Convention art. 20(2); Bern Convention art. 18(2).

[206] *See, e.g.*, IAEA Statute art. XIX.

[207] *See, e.g.*, Biological Weapons Convention art. VI **[fix Bluebook cite or do so above to full treaty name].**

With this wide range of possibilities for GOHAI's enforcement mechanisms, treaty drafters should aim to ensure certain desirable traits are included in these mechanisms. For instance, given the specialized knowledge needed to understand the functioning of high-risk AI, treaty drafters should seek to ensure that judges, arbitrators, or decision-makers possess a competent level of scientific literacy. Also, treaty drafters should ensure that binding dispute resolution mechanisms may be invoked by at least one party to a dispute and not permit parties to evade such binding mechanisms.[208] Finally, if more than one enforcement mechanism is included in the treaty, drafters should aim to ensure that all methods provided are effective and part of a wider process commencing with methods for non-combative resolution.[209]

2. Secretariat and depositary

As the permanent administrative department of an international body, a Secretariat will be needed for GOHAI. Secretariat functions may include organizing meetings of Parties, preparing reports and annual status reports, and conducting scientific and technical studies.[210] GOHAI's Secretariat should be further empowered to provide Parties with information on private entities that can provide technical assistance on high-risk AI, especially considering the concentration of AI talent in the private sector.[211] Also, a catchall provision permitting the Secretariat to complete any function assigned to it by the Parties would be useful.[212] Finally, treaty drafters should consider whether it is desirable to establish a stand-alone Secretariat or to request the services of the United Nations Secretary-General as the treaty Secretariat.[213]

Further, the same entity that serves as the Secretariat should serve as the Depositary, which will be entrusted with the treaty itself.[214] A Depositary's functions may include informing Parties of the date when a treaty will come into force, as well as providing notifications of any withdrawals, amendments, and updates to any annexes.[215]

3. Financial mechanisms

For the financing of GOHAI and its initiatives, treaty drafters should seek to implement mechanisms that will facilitate predictable, robust funding options. Looking to existing practices, treaties integrated into the United Nations' bodies may delegate determination of expenses to the

[208] *See, e.g.*, CITES art. XVIII (requiring mutual consent of *both* Parties for arbitration).
[209] *See, e.g.*, VCPOL art. 11 (allowing negotiation, arbitration, referral to the ICJ, or a conciliation commission).
[210] Pervaze A. Sheikh & M. Lynne Corn, Cong. Rsch. Serv., RL32751, *The Convention on International Trade in International Trade in Endangered Species of Wild Fauna and Flora (CITES)*, at 6 (2016) [hereinafter *CITES CRS Report*]. For other descriptions of the functions of Secretariats, see SCND arts. 16, 18; VCPOL art. 7; Montreal Protocol arts. 7, 12; CBD art. 24; Cartagena Protocol art. 31; CITES art. XII; Basel Convention art. 16.
[211] *See, e.g.*, Basel Convention art. 16(h).
[212] *See, e.g.*, CITES art. XII(2)(i).
[213] *See, e.g.*, SCND art. 16.
[214] *See, e.g.*, CBD art. 41; Basel Convention art. 28; VCPOL art. 20.
[215] *See, e.g.*, VCPOL art. 20.

United Nations General Assembly.[216] Similarly, Parties may have a subgroup of the Parties propose an initial annual budget that must be subsequently approved by a two-thirds majority of all Parties to a treaty.[217] To supplement these budgetary mechanisms, treaty bodies may also establish trust funds to finance treaty initiatives.[218]

Also, financial mechanisms may also be dedicated to supporting the spread of information or low-income states' enforcement of the treaty's objectives. For example, the Montreal Protocol established a Multilateral Fund that supports developing countries and clearing-house functions of spreading useful information.[219] For GOHAI, such a mechanism could assist low-income countries in establishing their own specialized bodies of AI experts.

4. Individual Parties' specialized bodies of experts

Given the need to oversee high-risk AI under the control of private entities and the need for focal points of coordination with GOHAI, each state Party to GOHAI should be required to establish designated Management and Scientific Authorities.[220]

For instance, CITES requires all Parties to designate a Management and Scientific Authority for its objectives.[221] Through regulatory tools, including import and export controls, CITES seeks to protect endangered animal and plant species from overexploitation because of international trade.[222] According to Article IX of CITES, each Party must designate a Management Authority "competent to grant permits or certificates on behalf of that Party."[223] Article IX also requires each Party to establish a Scientific Authority, which, among other responsibilities, must determine whether or not the granting of an import or export permit will have a harmful effect on the conservation status of a species.[224]

Similarly, based on the CITES system, GOHAI could mandate a global network of Management and Scientific Authorities. Under one approach, GOHAI could oversee in part a global licensing system that has each state Party outlaw certain high-risk AIs by default and only permit their release or training with a license.[225] Under this framework, each state Party's

---

[216] *See, e.g.*, VCPS art. 24.
[217] *See, e.g.*, IAEA Statute art. XIV.
[218] *See, e.g.*, Decisions of the Meetings of the Parties to the Montreal Protocol, Decision I/14: Financial arrangements, https://ozone.unep.org/sites/default/files/Handbooks/MP-Handbook-2020-English.pdf, at 650. **[fix Bluebook]**
[219] *See, e.g.*, Montreal Protocol art. 10.
[220] *See, e.g.*, VCPS art. 6; SCND art. 17; CITES art. IX; Basel Convention art. 5.
[221] In the United States, the Office of Management Authority (OMA) and the Office of Scientific Authority (OSA) of the Department of the Interior (DOI) exercise these responsibilities. *CITES CRS Report* at 2.
[222] CITES, Preamble.
[223] CITES, art. IX(1)(a). For a concise summary of a Management Authority's responsibilities, see *CITES CRS Report* at 6.
[224] CITES, art. IX(1)(b); *see also CITES CRS Report* at 6.
[225] In the United States, however, such a system that outlaws the release of high-risk foundation models may encounter challenges on First Amendment free speech grounds. *See Bernstein v. U.S. Dep't of Just.*, 176 F.3d 1132 (9th Cir.), *reh'g granted*, *op. withdrawn*, 192 F.3d 1308 (9th Cir. 1999); *see also* Xiangnong Wang, *De-Coding Free Speech: A First Amendment Theory for the Digital Age*, 2021 Wis. L. Rev. 1373 (2021).

Scientific Authority would assess the likelihood of harm from a given model before its release, as well as the likelihood of compliance with global safety standards. Based partially on the Scientific Authority's assessment, which should be determinative as to scientific matters, the Management Authority could then determine whether or not to grant a license to a model or to impose further conditions on its training or release, determining as well if further communication with the Secretariat or other Parties is needed regarding that model.

5. A global impartial body of experts

To help facilitate public trust in technical expertise in relation to AI, a global impartial, apolitical body of experts should be established as a separate body from GOHAI's policymaking and advocacy entities. For example, the Single Convention on Narcotic Drugs (SCND), the Vienna Convention on Psychotropic Substances (VCPS), and the Convention Against Illicit Traffic in Narcotic Drugs and Psychotropic Substances (CAIT) (together, the "Drug Treaties") partially rely on the World Health Organization (WHO) as a stand-alone body for some of their core functions.[226] As a United Nations agency, WHO aims "to promote health, keep the world safe and serve the vulnerable–so everyone, everywhere can attain the highest level of health."[227] To ensure the safety of the world from AI threats, a new United Nations agency, the World Artificial Intelligence Organization (WAIO) could be established.

Alternatively, the treaty creating GOHAI could create a subsidiary body that would serve this similar function and be composed of a global, impartial group of experts. For example, the Convention on Biological Diversity (CBD) and the Cartagena Protocol on Biosafety to the Convention on Biological Diversity (the Cartagena Protocol) rely on a Subsidiary Body on Scientific, Technical and Technological Advice (SBSTTA) to provide state Parties with advice relating to the implementation of the CBD and the Cartagena Protocol.[228] However, the SBSTTA meets on a more intermittent basis than the more permanent WHO.[229] Given the significant threats posed by AI, treaty drafters should favor the creation of a more permanent organization like WHO.

Finally, to further cooperation for AI safety research, WAIO could serve a clearing-house function as well to facilitate the exchange of information, assist with implementation of the treaty, and make non-dangerous information freely available.[230] However, further research should be conducted to determine the best manners of efficiently approaching AI safety research, considering as well existing initiatives, such as the Global Partnership on Artificial

[226] *See, e.g.*, SCND art. 3(1); VCPS art. 2(1); and CAIT art. 14(4).
[227] About WHO, World Health Organization, https://www.who.int/about.
[228] CBD art. 25.
[229] *See* Convention on Biological Diversity, *Subsidiary Body on Scientific Technical and Technological Advice (SBSTTA)*, https://www.cbd.int/sbstta/#:~:text=Article%2025%20of%20the%20Convention,other%20subsidiary%20bodies%2C%20with%20timely.
[230] Cartagena Protocol art. 20; *see also* Biosafety Clearing-House, https://bch.cbd.int/en/.

Intelligence.[231] Treaties may further establish scientific centers for research and education relating to their treaty objective.[232] Drafters should also consider methods of integrating AI experts working in the private sector in such initiatives, considering the concentration of AI talent in the private sector.

6. A global body of experts with the mandate to advocate for the execution of the treaty

Apart from the impartial body of experts operating under WAIO, GOHAI should have a separate body of experts with the express mandate to advocate for the success of the treaty's objectives. With AI's complicated nature and the difficulty of driving collective action from many state Parties, an advocacy-focused group of experts would help increase GOHAI's effectiveness. Further, separating WAIO and the advocacy-focused body of experts will help preserve public trust in WAIO's impartialness.

For example, the Drug Treaties entrust a variety of functions to the International Narcotics Control Board (the "Board"), empowering the select group of experts on the Board to pursue measures that will heighten the effectiveness of the treaties' objectives.[233] The Board is composed of thirteen members elected by the Economic and Social Council of the United Nations (ECOSOC), and Board members serve terms of five years, subject to removal by ECOSOC under certain conditions.[234] To be clear, the Board is tasked with many tasks that hinge on impartiality and its credibility, including administering an estimate system of drug requirements;[235] administering the statistical returns system of various uses of regulated drugs and related items;[236] and preparing reports on its work.[237]

However, the Drug Treaties also task the Board with key advocacy functions aiming to ensure the treaties' effectiveness.[238] For instance, in situations where a state has become or risks becoming a center for the illicit cultivation, production, manufacture, trafficking, or consumption of regulated drugs, SCND provides that the Board may pursue a process of consultations with the state; request remedial measures; propose and offer assistance in a study of the matter; and, if other measures fail, call the attention of all Parties and the global public to the matter.[239]

Like the Board operating under the Drug Treaties, GOHAI could establish an International High-Risk AI Control Board ("AI Control Board") to promote execution of the

---

[231] The Global Partnership on Artificial Intelligence (GPAI), About GPAI, https://gpai.ai/about/. Many articles on AI safety are also published on arxiv.org. *See* note 7.
[232] *See, e.g.*, SCND art. 38 bis.
[233] *See, e.g.*, SCND art. 14; VCPS art. 19; CAIT art. 22.
[234] SCND arts. 9–10.
[235] SCND arts. 12, 19.
[236] SCND arts. 13, 20.
[237] SCND art. 15; VCPS art. 18. The SCND also states Board members should possess "impartiality and disinterestedness." SCND art. 9(2).
[238] SCND arts. 14; VCPS art. 19; CAIT art. 22. *See also* SCND art. 14 bis (permitting the Board to recommend technical or financial assistance to a state Party, as long as said Party agrees).
[239] SCND art. 14. This process also requires the concern to arise "without any [State] failure in implementing the provisions of the [SCND]." SCND art. 14(1)(a). *See also* VCPS art. 19; CAIT art. 22.

treaty's objectives. Under a similar process as that tasked to the International Narcotics Board by the Drug Treaties, the AI Control Board could work with individual state Parties and draw the attention of state Parties and the international community to facilitate safety in the training or release of prominent high-risk models. Similarly, if a state party has become or risks becoming a center for the illicit training or production of high-risk models, the AI Control Board could work with that state party or draw the attention of the global community as needed.

7. Representative entities that oversee treaty implementation, including updating a given high-risk model's risk tier

GOHAI should also include entities that represent state Parties and are empowered to oversee implementation of the treaty, including in updating AIs' risk-tier classifications.

For example, the Drug Treaties empower the Commission on Narcotic Drugs (the "Commission") and the Economic and Social Council of the United Nations (ECOSOC) to update the risk tier classifications of substances and their inputs.[240] ECOSOC is one of the six main organs of the United Nations (UN) and aims to advance three areas of sustainable development: economic, social, and environmental initiatives.[241] ECOSOC is composed of 54 state Parties "elected for three-year terms by the [UN] General Assembly."[242] Subject to oversight by ECOSOC, the Commission is the governing body of the United Nations Office on Drugs and Crime (UNODC) and oversees many regulatory activities detailed in the Drug Treaties.[243] The Commission is composed of 53 state Parties elected by ECOSOC and "is chaired by a Bureau, including one member per Regional Group."[244]

Assuming integration with UN entities is desired, treaty drafters could create a Commission on High-Risk Artificial Intelligence (CHAI) that is similarly empowered as the Commission on Narcotic Drugs, while remaining subsidiary and subject to the control of ECOSOC. Like the Commission, CHAI could play a leading role in updating the risk tier classifications of regulated substances and inputs: namely, AI models and ML chips.[245]

Alternatively, GOHAI could mimic the structure of the International Atomic Energy Agency's (IAEA) policymaking bodies, particularly the General Conference and the Board of Governors. The General Conference consists of all IAEA state Parties and has various powers, including electing members of the Board of Governors, suspending Parties, and approving amendments to the IAEA Statute.[246] The Board of Governors consists of 35 state Parties that are geographically representative of the General Conference and also representative of the "most

[240] SCND arts. 7–8; VCPS art. 17; CAIT art. 21.
[241] United Nations Economic and Social Council, About Us, https://www.un.org/ecosoc/en/content/about-us.
[242] United Nations Economic and Social Council, FAQ, https://www.un.org/ecosoc/en/FAQ.
[243] United Nations Commission on Narcotic Drugs, https://www.unodc.org/unodc/en/commissions/CND/index.html.
[244] *Id.*
[245] *See supra* note 241.
[246] IAEA Statute arts. V(A), V(E)(1), V(E)(3), and V(E)(9). *See also* IAEA General Conference, About Us, https://www.iaea.org/about/governance/general-conference.

advanced [members] in the technology of atomic energy."[247] The Board of Governors has a variety of powers, including recommending to the General Conference a budget for the IAEA; considering applications for memberships; approving safeguards agreements with individual state Parties; approving the publication of the IAEA's safety standards; and appointing the Director General of the IAEA, with the approval of the General Conference.[248] To make decisions, both the General Conference and the Board of Governors generally require a two-thirds majority vote.[249]

Applying the IAEA model to global policymaking entities for high-risk AI, the more representative entity (General Conference) could be delegated similar functions as the Commission, and the governing body (the Board of Governors) could have similar powers as ECOSOC.

8. A process for updating high-risk models' risk tier classifications

Given high-risk AIs' potential to have emergent capabilities or to pose previously undiscovered risks,[250] GOHAI should have a relatively rapid process for updating high-risk AIs' and their inputs' risk tier classifications.

The Drug Treaties offer a potential model for such a process, specifically their process for updating the risk tier classifications of both regulated drugs and their inputs; this process also applies to adding or removing regulated substances. Using a risk-based approach to impose varying protective safety measures on regulated drugs and their inputs, the Drug Treaties provide a comprehensive classification scheme. CAIT divides inputs, or precursor substances capable of becoming regulated drugs into Table I and Table II, subjecting those precursor substances to different requirements based on their respective Table.[251] SCND and VCPS divide the regulated drugs themselves into Schedules I, II, III, and IV, subjecting drugs to different requirements based on their respective Schedule.[252]

For inputs, CAIT article 12 describes the process for whether an input is moved between Table I or II, is added, or is removed. First, a state Party or the International Narcotics Control Board (the "Board") may provide information to the UN Secretary-General of a possible needed change for a specific input's classification.[253] Second, the Secretary-General subsequently communicates this information to all state Parties, the Commission on Narcotic Drugs (the

[247] IAEA Statute art. VI(A); *see also* IAEA Board of Governors, About Us, https://www.iaea.org/about/governance/board-of-governors#:~:text=The%2035%20Board%20Members%20for,the%20Russian%20Federation%2C%20Saudi%20Arabia%2C.
[248] IAEA Board of Governors, About Us, https://www.iaea.org/about/governance/board-of-governors#:~:text=The%2035%20Board%20Members%20for,the%20Russian%20Federation%2C%20Saudi%20Arabia%2C.
[249] IAEA Statute arts. V(C), VI(E).
[250] *See supra* footnote 88.
[251] CAIT art. 1(t).
[252] SCND art. 1(u); VCPS arts. 1(e), (g).
[253] CAIT art. 12(2).

"Commission"), and the Board, collecting further additional information as provided.[254] Third, based on this information, the Board decides whether the input is (a) "frequently used in the illicit manufacture of a narcotic drug or psychotropic substance"; and whether (b) "the volume and extent of the illicit manufacture of a narcotic drug or psychotropic substance creates serious public health or social problems . . . ."[255] The Board's assessment is "determinative as to scientific matters."[256]

Fourth, based on comments of state Parties and the Board's determination, the Commission decides on a two-thirds majority vote as to whether or not to enact the change to the input's classification.[257] Fifth, the Secretary-General communicates the Commission's decision to all state Parties and entities, and the Commission's decision is effective upon 180 days after this communication.[258] Sixth, if a state Party disagrees with this classification, it may appeal the Commission's decision to ECOSOC, which may confirm or reverse the decision.[259]

For the regulated drugs themselves, the SCND and VCPS offer a similar process for changes to the risk classification, addition, or deletion of a regulated drug to or from one of the four Schedules.[260] While the process is similar to that for inputs under CAIT, there are a few differences. For example, in the first step, the SCND and VCPS permit any state Party or WHO to notify the Secretary-General of a possible needed change, but not the Board.[261] In the second step, while all state Parties and the Commission receive communications, the WHO replaces the Board in receiving communications from the Secretary-General.[262] Further, pending the final decision of the Commission as to the substance's classification, SCND permits the Commission to provisionally consider that substance as a Schedule I substance–the highest risk tier–and state Parties are required to apply control measures provisionally to this substance.[263]

Third, instead of the Board making a determination as for inputs, the WHO is given the task of judging the potential harm of a substance.[264] Fourth, based on the WHO's communications and any other relevant information, the Commission decides on whether or not to change a substance's classification.[265] Fifth, while not an option under SCND, VCPS permits state Parties to opt out of certain aspects of the new classification, subject to ongoing minimal

---

[254] CAIT art. 12(3).
[255] CAIT art. 12(4). In making this determination, the Board must also consider (i) the "extent, importance and diversity of the licit use of the substance" and (ii) the "possibility and ease of using alternate substances both for licit purposes" and illicit purposes. *Id.*
[256] CAIT art. 12(5).
[257] CAIT art. 12(5).
[258] CAIT art. 12(6).
[259] CAIT art. 12(7).
[260] *See* SCND art. 3; VCPS art. 2.
[261] *Compare* SCND art. 3(1); VCPS art. 2(1) *with* CAIT art. 12(2).
[262] *Compare* SCND art. 3(2); VCPS art. 2(2) *with* CAIT art. 12(3).
[263] SCND art. 3(3)(ii). The VCPS takes a less binding approach: state Parties may "examine . . . the possibility" of such a provisional re-classification of a substance. VCPS art. 2(3).
[264] SCND arts. 3(3)(iii), 3(4), 3(5); VCPS art. 2(4). For psychotropic drugs, the WHO's determinations are "determinative as to medical and scientific matters." VCPS art. 2(5).
[265] SCND art. 3(6); VCPS arts. 2(5), 2(6).

obligations for control measures.[266] Sixth, SCND permits state Parties to appeal this new classification to ECOSOC within 90 days of notification from the Secretary-General of the decision, while VCPS permits this appeal up to 180 days from said notification.[267] Under both SCND and VCPS, ECOSOC may then "confirm, alter or reverse" the Commission's decision, and such determination will be final.[268]

For high-risk AI, the Drug Treaties' process for risk classification, addition, or deletion of a regulated substance and its inputs may be a viable path forward for international regulation. Instead of two classification processes for precursor substances and regulated drugs, GOHAI could have two classification processes for ML chips (inputs) and high-risk AI models themselves. This proposal assumes two tiers for ML chips' classification and at least two tiers of risk classification for AI models.

First, for both processes, any state Party or its Management and Scientific Authorities, WAIO, or the AI Control Board should be empowered to notify or provide relevant information to the Secretariat, regarding a high-risk model.[269] Among these entities, I do not see any compelling reason to limit which entity may inform the Secretariat of such a concern; more scrutiny of dangerous models is desirable, and these entities are unlikely to share frivolous concerns.

Second, with the input of the AI Control Board, the Secretariat could determine what information should be communicated to which relevant parties,[270] considering as well the possibility that some information should not be widely communicated. For instance, if the AI Control Board is made aware of a high-risk model capable of designing bioweapons, the Secretariat should not publicize that information and merely communicate with those entities that *need* to know that information to mitigate that risk, such as the AI's developer, state of origin, and possibly states exposed to harm by the model.

Third, for ML chips' classification, the AI Control Board could make a determination as to ML chips' capabilities to inform which risk tier should apply to a given ML chip.[271] For instance, given that ML chips are dual-use hardware, a future regulation could place limits on the use of certain ML chips, distinguishing their use for purposes, such as climate modeling, as opposed to building high-risk AIs.[272] Similarly, for high-risk models' classification, WAIO could conduct a risk assessment as to the models' capabilities, a determination that should be determinative as to scientific and technological matters.[273] Pending the AI Control Board's or WAIO's decisions, the Commission on High-Risk AI (CHAI) should be empowered to

[266] VCPS art. 2(7).
[267] SCND art. 3(8); VCPS art. 2(8).
[268] SCND art. 3(8)(c); VCPS art. 2(8)(c).
[269] *See* SCND art. 3(1); VCPS art. 2(1); CAIT art. 12(2).
[270] *See* SCND art. 3(2); VCPS art. 2(2); CAIT art. 12(3).
[271] *See* CAIT art. 12(4).
[272] *See* Shavit, *Compute Monitoring*, *supra* note 10, at 9.
[273] *See* VCPS art. 2(5).

provisionally classify a ML chip or AI model into a given Schedule, and this classification should be binding on state Parties to apply the tier's respective safety measures.[274]

Fourth, for both processes, CHAI should be empowered to make an informed decision on the classification of a ML chip or high-risk AI model. Fifth, the Secretariat should communicate this decision to all entities. Sixth, given the need for rapid responses to high-risk AIs, the classification should be applied provisionally by the state Parties and take permanent effect as soon as is reasonably possible, such as after 90 days.[275] Seventh, within a reasonable, short time-frame, state Parties should be permitted to appeal CHAI's decision to ECOSOC, which should be enabled to reverse, alter, or confirm CHAI's decision.[276]

This proposed system may need further modifications, and I encourage further research on processes that may permit more rapid re-classifications, that account for the prevalence of top AI talent in the private sector, or that consider manners of accomplishing similar safety objectives from a domestic law approach.

9. A representative body of all parties, including possible subsidiary bodies

Finally, meetings of the Parties should be arranged on a consistent, frequent basis, possessing the ability to rapidly convene in cases of emergency developments. While I have noted possible entity structures for a treaty integrated into the UN, the Conference of the Parties (COP) model is an alternative approach for treaty bodies not fully integrated in the UN.[277] Possible functions of the COP include the ability to vote on amendments and to review the effectiveness of the treaty.[278]

III. Global Governance for Accountability

Part II engaged in the exercise of defining existential risks from unaligned, high-risk AI, concluding that an objective of promoting, maintaining, and enforcing safety measures would partially address this risk. Part III engages in a similar exercise, outlining existential risks from misuse of high-risk AI and arguing for an objective of maintaining accountability for state Parties and developers of high-risk AI to address such misuse risks. After presenting existential risks from high-risk AI misuse, this Part explores possible precedent legal models for such mechanisms, suggesting possible adaptations for a system that would enable accountability of states and actors working with high-risk AI.

---

[274] *See* SCND art. 3(3)(ii).
[275] *See* SCND art. 3(8).
[276] SCND art. 3(8)(c); VCPS art. 2(8)(c). The COP model may also be compared to the IAEA's General Conference. *See* IAEA Statute art. V.
[277] VCPOL art. 6; Convention on Biological Diversity art. 23; CITES art. XI; Basel Convention art. 15.
[278] *CITES CRS Report* at 7.

By accountability, I refer to mechanisms that enable transparency, oversight, complaints, and enforcement of the treaty's objectives.[279] By *transparency*, I mean information-sharing or reporting mechanisms imposed on state Parties that require them to share information with the international community, other states, or particular entities within an international organization.[280] By *oversight*, I refer both to mechanisms that enable active monitoring of regulated substances under a treaty and of state Parties' compliance with treaty obligations.[281] *Complaint* mechanisms permit "grievances attributable to a country to be processed and adjudicated through country or secretariat complaint mechanisms."[282] Finally, while I have previously used the Hoffman definition for *enforcement*,[283] in this Part, I focus solely on enforcement mechanisms that enable the possibility of sanctions or consequences delivered by a court or other authority on a state Party for a failure to control private entities under its jurisdiction that have acted against a treaty's objectives.

While the Hoffman study found that transparency, complaint, and oversight mechanisms were not necessarily associated with a treaty being effective,[284] I argue such mechanisms may still be useful for global governance of high-risk AI. Specifically, given the nature of AIs as software that is extremely difficult to attribute to any individual or organization once released, mandating globally applicable transparency mechanisms will help ensure regulators are knowledgeable about the activities of states and developers of high-risk AI. For example, reporting requirements for certain high-risk AIs' training would ensure national regulatory authorities are able to hold organizations accountable for failing to comply with regulations. Similarly, assuming globally applicable safety standards or practices are implemented, oversight and complaint mechanisms would be necessary to permit the eventual use of an enforcement mechanism that would hold a state Party responsible for a failure to control a private entity. In other words, for high-risk AI, transparency, complaint, and oversight mechanisms may be necessary to ensure that enforcement mechanisms can actually be applied. Lastly, my definitions of *transparency*, *oversight*, and *enforcement* vary slightly from those used in the Hoffman study, so the study results may not be applicable for the used definitions.

A. Misuse of High-Risk AI

[279] Steven J. Hoffman et al., *International treaties have mostly failed to produce their intended effects* (Aug. 1, 2022), https://www.pnas.org/doi/full/10.1073/pnas.2122854119#t02.
[280] This definition is slightly more inclusive than that in the Hoffman study, which defines transparency mechanisms as those that "enable information to be shared about countries with observers through regular reporting and information aggregation." *Id.*
[281] This definition is also more inclusive than that in the Hoffman study, which defines oversight mechanisms as those that "build on transparency by actively monitoring and evaluating countries through standard setting and implementation review . . . ." *Id.*
[282] *Id.*
[283] *See supra* note __ (noting the possibility of a specific sanction or consequence by a court or other legal authority).
[284] Steven J. Hoffman et al., *International treaties have mostly failed to produce their intended effects* (Aug. 1, 2022), https://www.pnas.org/doi/full/10.1073/pnas.2122854119#t02 (referring to such mechanisms use in non-trade and finance treaties).

This section explains ways that high-risk AI may be misused by malicious or reckless actors. It argues these misuse cases pose existential risks to humanity and necessitate a global governance regime that would promote accountability, enabling further regulation or action if needed. In particular, it notes the extreme risks posed by high-risk AI capable of widely disseminating the means to create biological weapons, as well as AIs that autonomously seek to kill billions of humans.

First, misuse of increasingly capable and accessible high-risk AI enhances the risk of bioterrorism. For instance, foundation models, such as the systems behind ChatGPT that are rapidly improving, may enable non-expert users to create deadly known pathogens with step-by-step instructions.[285] These risks are enhanced by the potential diffusion of similar foundation models, a scenario that would increase the amount of people capable of creating such pathogens from a current number of about 30,000–120,000 experts to any non-expert with access to such a model.[286] Along with an AI's ability to enhance the probability of the creation of a known pathogen, current AIs are already capable of designing novel biological or chemical weapons.[287]

In a prominent example of this danger, a pharmaceutical company's research team discovered its AI designed to penalize toxicity and bioactivity could easily be repurposed to reward these traits, creating a tool for the design of new biological weapons.[288] After being trained on a public database's molecules, the repurposed model produced new molecules that were more toxic than previously known chemical warfare agents–despite the fact that none of the datasets used for training included those nerve agents or even the same "region of molecular property space" as the molecules in the previous model.[289] This area is poorly regulated, and researchers are largely in control of whether or not they choose to proceed with such dangerous experiments.[290]

Second, along with AI-related bioterrorism risks, an additional misuse case would involve a human willingly or recklessly developing or releasing unaligned AIs or AIs tailored to kill. Given the likelihood of AGI being unaligned,[291] its intentional or accidental release could result in negative results for humanity. Problematically, some leading AI leaders are declared "accelerationists," believing AIs are destined to replace humans as a dominant species and that humans should nevertheless strive to create AGIs.[292]

---

[285] Dan Hendrycks, Mantas Mazeika, & Thomas Woodside, *An Overview of Catastrophic AI Risks* (July 11, 2023), at 7, https://arxiv.org/pdf/2306.12001.pdf.
[286] Kevin M. Esvelt, *Delay, Detect, Defend: Preparing for a Future in which Thousands Can Release New Pandemics* (November 2022), at 11, Geneva Paper 29/22, https://dam.gcsp.ch/files/doc/gcsp-geneva-paper-29-22.
[287] Hendrycks et al., *An Overview of Catastrophic AI Risks*, *supra*, note 63, at 7.
[288] Fabio Urbina et al., *Dual use of artificial-intelligence-powered drug discovery*, 4 Nature Machine Intelligence 189 (2022).
[289] *Id.* at 189–90.
[290] *Id.* at 190.
[291] *See* Part II.A.
[292] Hendrycks et al., *An Overview of Catastrophic AI Risks*, *supra*, note 63, at 7 (describing accelerationist statements from Google co-founder Larry Page and eminent AI scientists Jürgen Schmidhuber and Richard Sutton). I am not commenting on whether future AGIs should have rights; rather, I aim to emphasize that accelerationist

Also, foundation models, such as those behind ChatGPT, can be repurposed for dangerous objectives.[293] For instance, an anonymous programmer took advantage of an open-source project to bypass ChatGPT's safety filters,[294] creating Chaos-GPT, an autonomous AI with goals to "destroy humanity; establish global dominance; cause chaos and destruction; control humanity through manipulation; and attain immortality."[295] Fortunately, ChaosGPT was nowhere near competent enough to accomplish its goals, but as AI progresses, the likelihood that a future AI would succeed with harmful goals increases.[296]

Given the unique characteristics of high-risk AI, accountability of key actors is needed to mitigate harmful results. At its core, a foundation model is software. Once this software is released to the general public, it is nearly impossible to recall. Further, as demonstrated by ChaosGPT, attributing liability to an actor intent on using released AIs for harm is difficult. Thus, one of the most effective areas of regulation may be targeting the entities developing such models or the activities used to create high-risk models. Further, after a high-risk model is created, regulation is critical to ensure nefarious actors do not fine-tune or model prompt it to create a future, more capable ChaosGPT, especially given that the resources to create an initial, raw foundation model are more scarce than those needed for fine-tuning or model prompting.[297]

While domestic regulation is also crucial for ensuring accountability for developers of foundation models, international regulation is a necessary complement, considering the effects of AI misuse would likely impact all states.[298] An international regime resulting in accountability of key actors–both private entities and states–and actions building toward high-risk AI will also enable emergency responses or regulation when necessary.

B. Accountability Mechanisms for High-Risk AI

This section explores accountability mechanisms that would enable transparency, oversight, complaints, and enforcement of high-risk AI. It begins with a preliminary overview in **Table 2** of such accountability mechanisms in existing international regulatory regimes. It then provides an in-depth overview of possible model regimes for accountability mechanisms,

---

viewpoints must recognize that a transition of control from humans to unaligned AGIs would likely involve enormous amounts of suffering.

[293] Apart from fine-tuning, model prompting permits persons to use different prompts to alter model behavior, instead of changing a model's parameters through fine-tuning. *See, e.g.*, Andy Zou et al., *Universal and Transferable Adversarial Attacks on Aligned Language Models*, ArXiv.org (July 27, 2023), https://arxiv.org/pdf/2307.15043.pdf.

[294] In other words, the model behind ChatGPT was unaligned, *and* the safety filters meant to mitigate against the model's flawed outputs failed.

[295] Hendrycks et al., *An Overview of Catastrophic AI Risks*, *supra*, note 63, at 8. Jose Antonio Lanz, *Meet Chaos-GPT: An AI Tool That Seeks to Destroy Humanity* (Apr. 13, 2023), Decrypt.co, https://decrypt.co/126122/meet-chaos-gpt-ai-tool-destroy-humanity.

[296] Hendrycks et al., *An Overview of Catastrophic AI Risks*, *supra*, note 63, at 8.

[297] The initial creation of a foundation model requires scarce ML chips, large amounts of compute for training runs, and large investments, while fine-tuning through supervised learning and RLHF can be done with many willing humans and lesser amounts of compute. *See* Part I.A.

[298] Shavit, *Compute Monitoring*.

applying relevant components to a global governance regime for mitigating misuse of high-risk AI.

Table 3: Accountability Mechanisms in Existing Global Legal Regimes

	Transparency mechanisms (*E.g.*, reporting or disclosure obligations)	Oversight mechanisms	Complaint mechanisms	Enforcement mechanisms for private entities
SCND	Yes, arts. 12(3), 13(1), 18, 19, 20, 24(2), 27(2), 35(f), 35(g)	Yes, arts. 15(1), 21, 21 bis, 22, 24, 25, 26, 28, 29, 30, 31, 32, **34**	No	Yes, arts. 4(c), 35(a), **36**
VCPS	Yes, arts. 2(1), 3(4), 12(2)(c), 13, **16**	Yes, arts. 5(2), 7(c), 8, 10, 11, 12, 14, 15	No	Yes, arts. 21(a), **22**
CAIT	Yes, arts. 12(9)(c), 12(10)(a), 12(11), 12(12), **20**	Yes, arts. 9(1), 9(3), 12(9)(a), 16, **22**	No	Yes, **arts. 3**, **4**, 5, **6, 7**, 15
IAEA Statute	Yes, arts. VIII, IX, X	Yes, arts. III(A)(5), IX, X, **XII**	No	No
The Structure and Content of Agreements Between the Agency and States Required in Connection with the Treaty on the Non-proliferation of Nuclear Weapons (**INFCIRC/153 (Corr.)**)[299]	Yes, paragraphs 8; 33–34; 59–69.	Yes, paragraphs 1–3; 7; 18–19; 70–82.	Yes, paragraph 9.	No, but arbitration for lack of state compliance may result in accountability for private actors. *See* paragraphs 20–22.

[299] **[Bluebook Cite]** [Mention what this doc is (how it is only guidance and how it is not binding) and how it is only applicable guidance for non-nuclear weapons states].

Montreal Protocol	Yes, arts. 2(5), 2(7), 7	Yes, arts. 4(1), 4(3), 4(6)	No	No
Convention on Biological Diversity (CBD)	Yes, arts. 14(1)(d), 17, 26; annex I	Yes, art. 7 (self-monitoring)	No	No, *but see* art. 27 & annex II for settlement of disputes among state Parties.
Cartagena Protocol to CBD	Yes, arts. 12, 14(2), **17(1), art. 20(3),** 25(3), 33	Yes, arts. 8–10, 13, 21, annex I	No	Yes, arts. 25(1), 27
Nagoya-Kuala Lumpur Supplementary Protocol on Liability and Redress to the Cartagena Protocol	No	No	No	Yes, arts. 5, 12
BWC	No	No	Yes, art. VI.	No
CITES	Yes, art. VIII(6)–(8)	Yes, arts. III, IV, V, VI, XIII (Secretariat monitoring)	No	Yes, art. VIII(1)–(2)

1. Transparency mechanisms

This section explores possible transparency mechanisms that are commonly used in risk-based treaties or treaties governing dual-use substances. For each mechanism, it specifies manners in which it could be applied to govern high-risk AI.

First, many treaties mandate periodic reports on the workings, implementation, or effectiveness of the treaty and its objectives. For example, the Drug Treaties, CITES, the Convention on Biological Diversity (CBD), and the Cartagena Protocol mandate periodic or annual reports on the working of the treaties, including the text of all laws and regulations giving effect to or changing obligations in relation to those treaties.[300] Further, the IAEA Statute and the Cartagena Protocol mandate reports that share information relating to the effectiveness of

[300] SCND arts. 18(1)(a)–(b) (mandating reports be given to the UN Secretary-General); VCPS art. 16(1) (same and covering "[s]ignificant developments" relating to the treaty); CAIT art. 20 (same); CITES art. VIII(7)–(8) (same); CBD art. 26 (same); Cartagena Protocol art. 33 (same).

projects or initiatives created by their respective treaty entities, aiming to advance knowledge relating to the peaceful use of nuclear energy and to living modified organisms, respectively.[301]

In the context of global governance for high-risk AI, a foundational treaty could establish similar reporting requirements. For instance, state Parties to a treaty establishing GOHAI could be required to provide an annual report on the workings, implementation, and effectiveness of the treaty, specifying changes or new laws and regulations concerning matters covered in the treaty. Further, if the treaty provides for initiatives that result in the production of new knowledge–such as that relating to the alignment problem–state Parties could be encouraged or mandated to provide such helpful information in reports or correspondences with WAIO, CHAI, the International High-Risk AI Control Board, or an entity similar to the Biosafety Clearing-House established in the Cartagena Protocol.[302]

Second, to regulate possibly harmful substances, many treaties mandate that state Parties report breaches or possible breaches of a treaty's objectives or of state Parties' obligations. For instance, the Drug Treaties require reports of illicit trafficking of regulated drugs;[303] imports, exports, or transits of precursory substances that may result in the illicit manufacture of regulated substances;[304] and seizures of such precursory substances.[305] Similarly, the CBD and Cartagena Protocol encourage state Parties to notify the global community of risks to biological diversity arising from their jurisdiction, as well as to take measures to reduce the likelihood of any transboundary movements posing such risks.[306]

Depending on which AIs are regulated, a treaty for high-risk AI could mandate reporting of different possible harms or manifested harms against which the treaty seeks to avoid. For instance, narrow models trained to be used in specific contexts, such as for pharmaceutical research, pose the risk of enabling non-experts to design biological weapons.[307] In the event where a company learns that one of its models has been misused for such a purpose either by a researcher or a licensee of such a model, the treaty could require state Parties to mandate that companies under its jurisdiction report such incidents to national regulatory authorities, such as the recommended established Management and Scientific Authorities, as well as the WAIO and possibly WHO. These reports would need to consider methods to mitigate risks from such misuse, as well as manners of confidentially sharing such risks with authorities that could enact mitigation measures.

Third, given that a treaty for high-risk AI would likely oversee thousands of AI models or entities, it could require the maintenance of active records of both estimates and statistics of regulated substances in each risk tier, as well as related components to such substances. CITES mandates such a system, requiring state Parties to maintain records of the trade of endangered

[301] IAEA Statute Art. VIII; Cartagena Protocol art. 20(3).
[302] *See* Cartagena Protocol art. 20.
[303] SCND art. 18(1)(c); VCPS art. 16(3).
[304] CAIT art. 12(9)(c).
[305] CAIT art. 12(12).
[306] CBD art. 17; Cartagena Protocol arts. 17, 25.
[307] *See* Part III.A and footnote 287.

specimens included in Appendices I, II, and III.[308] The Drug Treaties require state Parties to provide the International Narcotics Control Board (the "Board") with various estimates[309] and statistics[310] relating to regulated drugs and related substances. Finally, with its quantitative approach to regulation, the Montreal Protocol requires initial reporting and annual reporting requirements of regulated substances both to establish a baseline regulatory amount and to compare progress against that baseline.[311]

To effectively regulate certain high-risk AI, Yonadav Shavit and Mauricio Baker have suggested systems that could be adapted to include reporting of certain materials or locations that could be used to create high-risk foundation models. Based partially on Shavit's suggested adaptations to ML chips that would enable impartial persons to determine the chips' past use in training runs,[312] state Parties could be required to report a baseline amount of non-compliant chips, as well as to annually report quantities for compliant chips, the number of high-quality ML chip fabrication facilities ("fabs"), data centers, storage facilities, and elimination facilities.[313] Importantly, this approach based on hardware and locations would be underinclusive for high-risk AI and would need to be reinforced by active records of other high-risk models, especially given that narrow AI for drug discovery or in other dangerous contexts poses high risks.

Fourth, in the context of import and export controls of regulated substances, many treaties require state Parties to report certain movements of such substances when crossing state Party borders. The Drug Treaties, for example, require or permit state Parties to notify other state Parties of exports of regulated drugs or precursory substances, as well as of prohibited imports of regulated drugs.[314]

For high-risk AI, given that the Biden administration has implemented trade controls to limit China's access to AI-related materials, a global treaty could impose similar reporting requirements when such materials are moved internationally, setting a foundation for a global oversight system.[315] With U.S. companies dominating many AI-related technologies, in October 2022, the Biden administration announced trade restrictions meant to severely limit China's access to high-end ML chips; ML chip-design software; semiconductor manufacturing equipment and maintenance resources; and components of semiconductor manufacturing

---

[308] CITES art. VIII(6).
[309] SCND art. 19 (requiring estimates for quantities relating to, among others, drugs used for medical and scientific purposes, for the manufacture of other drugs, and for the area of land to be used for opium poppy cultivation).
[310] SCND art. 20 (requiring statistics for quantities relating to, among others, production or manufacture of drugs, consumption of drugs, import and exports of drugs and poppy straw, and seizures of drugs); *see also* SCND art. 27 and VCPS art. 16(4)–(5).
[311] *See* Montreal Protocol art. 7.
[312] Shavit, *Compute Monitoring*.
[313] Mauricio Baker, *Nuclear Arms Control Verification and Lessons for AI Treaties* (April 8, 2023), arXiv, at 37–42, https://arxiv.org/pdf/2304.04123.pdf.
[314] VCPS arts. 12(2), 13; CAIT art. 12(10).
[315] For such a global AI oversight system, *see* Mauricio Baker, *Nuclear Arms Control Verification and Lessons for AI Treaties* (April 8, 2023), arXiv, at 37–42, https://arxiv.org/pdf/2304.04123.pdf.

equipment.[316] In the future, these trade restrictions could be expanded among U.S.-allied countries or on a global scale, requiring both reporting and oversight requirements to mitigate risks.

Fifth, with likely widespread diffusion of high-risk AI, state Parties could be requested to report possible needed regulation or changes to the risk-tier for a given AI model.[317]

2. Oversight mechanisms

This section explores possible models through which treaties enable oversight of regulated substances and of state Parties themselves. From these precedents, it then suggests possible oversight mechanisms for global governance of high-risk AI.

Generally, many treaties pursue oversight of regulated substances by having state Parties agree to take domestic measures that will be uniformly applied on a global scale, resulting in reduced risk of harm from such substances. For example, the VCPS requires state Parties to "maintain a system of inspection of manufacturers, exporters, importers, and wholesale and retail distributors of psychotropic substances and of medical and scientific institutions which use such substances."[318] The SCND encourages state Parties to require licensing of individuals working with regulated drugs, mandating such persons to have "adequate qualifications" to comply with laws and regulations enacted in compliance with the convention.[319] Further, requirements for record keeping by state Parties may be reinforced by requirements for record keeping by private entities.[320]

Global agreements for high-risk AI could encourage similar, uniform domestic measures that have been imposed in previous treaties. Like the SCND's encouragement of licensing individuals, national licenses could be required for any natural or legal person working with or in fabs, with high-quality ML chips, and in developing high-risk models, including foundation models and narrow, high-risk AI, such as those in the pharmaceutical context. Also, to have an effective state Party record-keeping system as explained in Part III.B.1, a state would need to require private entities to establish their own record-keeping systems of high-risk models, high-quality ML chips, and other components used to build such models.

Further, to supplement domestically-imposed measures, a treaty may seek to mitigate harms from a regulated substance by imposing oversight in the context of trade restrictions for transboundary movements.[321] As explained in Part II.E.4, CITES imposes a global system of

[316] Gregory C. Allen, *Choking off China's Access to the Future of AI* (Oct. 2022), Center for Strategic & International Studies, https://csis-website-prod.s3.amazonaws.com/s3fs-public/2023-04/221011_Allen_China_AccesstoAI.pdf?VersionId=V2RCmPjpR8nQOybvB2LmNIu1Yx6RlYvA.
[317] *See* VCPS arts. 2–3; *see also* the process suggested in Part II.E.8.
[318] VCPS art. 15.
[319] SCND art. 34(a).
[320] For such requirements for state Parties, see footnote 309.
[321] See Part III.B.1 for reporting mechanisms in the trade context.

import and export controls partially through reliance on state Parties' individual Management and Scientific Authorities to grant import or export permits and certificates.[322]

Based loosely on CITES, a global trade oversight system for materials used to train and build high-risk AIs could be effective in ensuring global compliance with shared objectives. While underinclusive as to high-risk AI, this system could focus on the materials targeted through the current Biden administration trade restrictions toward China.[323] In such a system, state Parties' Management and Scientific Authorities could oversee the granting of import and export permits and certificates; Scientific Authorities could determine compliance with Shavit's suggested requirements for ML chips; and Management Authorities could determine whether or not to permit the import or export of such ML chips.[324]

Finally, given the breadth of global oversight of nuclear-related materials and of parties dealing with such materials,[325] many have suggested the global nuclear regulatory system may offer guidance for regulation of high-risk AI, which will likely require similar robust oversight.[326] In making this comparison to a possible high-risk AI global regulatory regime, however, it is extremely important to note the objectives of the global nuclear regulatory regimes. Notably, the nuclear regimes have different, even conflicting objectives tailored to deal with different risks. Through international and national regulatory regimes, states seek to increase security of nuclear facilities to guard against possible misuse of or harm from unauthorized possession of nuclear materials;[327] to *increase* the proliferation of nuclear technology to guard against lack of reliable energy;[328] and to use safeguards to both *prevent* proliferation of nuclear weapons and to ensure safety from misuse of nuclear materials for the unauthorized creation of nuclear weapons.[329] In the nuclear regulatory context, *safeguards* are oversight activities through which the IAEA may verify a state is not using nuclear programs to create unauthorized nuclear weapons.[330]

This Article has thus far focused primarily on objectives relating to safety, not objectives similar to those in the nuclear context, such as proliferation of capable AIs or the prevention of unaligned AIs' proliferation. Even with this safety focus, however, comparisons to the nuclear safety regime may yield limited results. The IAEA focuses primarily on oversight of tangible

---

[322] See Part II.E.4; CITES art. IX.
[323] *See* footnote 315.
[324] *Compare with* Part II.E.4, where I recommend Management and Scientific Authorities could license AI models' use *within* a state, not in the trade context. Management Authorities of a state of export could also determine if the importing state is compliant with applicable global requirements for AI safety.
[325] Mauricio Baker, *Nuclear Arms Control Verification and Lessons for AI Treaties* (April 8, 2023), arXiv, at 25, https://arxiv.org/pdf/2304.04123.pdf.
[326] For an alternative method of minimal oversight of state Parties to a treaty, see CAIT art. 22.
[327] James M. Acton, *Governance of Dual-Use Technologies: Theory and Practice, Ch. 1: On the Regulation of Dual-Use Nuclear Technology*, https://www.amacad.org/publication/governance-dual-use-technologies-theory-and-practice/section/4 (the objective of security efforts is to "prevent the unauthorized possession of nuclear material . . . .").
[328] *See* IAEA Statute art. II.
[329] *See* IAEA Statute art. III.A.5.
[330] *IAEA Safeguards Overview: Comprehensive Safeguards Agreements and Additional Protocols*, IAEA, https://www.iaea.org/publications/factsheets/iaea-safeguards-overview.

materials used to create nuclear weapons; in the AI context, while many tangible materials are important for the creation of AI models, a finished model is ultimately just software. Also, with nuclear materials, there are two possible outcomes for their use–a state Party uses them for peaceful purposes, such as energy, or to create a nuclear weapon. For high-risk AI, use outcomes are not so clear-cut; an entity may attempt to use materials to make a capable, safe AI that is later discovered to be dangerous.[331]

Thus, rather than to focus on the possible implementation and workings of oversight mechanisms for high-risk AI based on the nuclear safety regime, I next briefly explore possible secondary objectives of similar oversight regimes for high-risk AI, drawing on established risks and recognizing that such a materials-based oversight system would be limited in its practicality.[332]

For global oversight of high-risk AI, one could reasonably frame prominent risks as those arising from misuse by an actor or from a highly capable, unaligned AI; from these risks, a high-priority objective would be ensuring safety from such risks.[333] Focusing on the former risk, misuse could be defined in various manners relative to non-compliance with specific restrictions, such as those based on total training compute, properties of training data, properties of hyperparameters, performance-based benchmarks, bans, or post-training safety mitigations.[334] Also, while not explored in this Article, security against nefarious actors seeking to steal highly capable model weights, as well as other sensitive information will likely be an important objective. States may also seek to limit or eliminate proliferation of ML chips that are not compliant with Shavit's suggested modifications. Regardless of the resulting objective, a monitoring system would require the establishment of a permanent staff for global inspections, possibly operating under GOHAI's direction.[335]

A final consideration from the global nuclear safety regime is the difference in treatment of nuclear-weapon states (NWSs) and non-nuclear weapon states (NNWSs). For instance, the NPT requires NNWSs to accept safeguards, while NWSs have no such binding obligation.[336] NNWSs largely accepted this two-tiered treatment in exchange for a disarmament commitment and a commitment for "'the fullest possible exchange' of nuclear materials, equipment, and knowledge between states."[337] While two separate groups may emerge in the context of

---

[331] *See* Part II.A.

[332] For such possible oversight mechanisms for high-risk AI and overviews of the agreements composing the global nuclear safety regime, see Shavit, *Compute Monitoring*; Mauricio Baker, *Nuclear Arms Control Verification and Lessons for AI Treaties* (April 8, 2023), arXiv, https://arxiv.org/pdf/2304.04123.pdf; Laura Rockwood, *Legal Framework for IAEA Safeguards*, IAEA, https://www.iaea.org/sites/default/files/16/12/legalframeworkforsafeguards.pdf; *Basics of IAEA Safeguards*, IAEA, https://www.iaea.org/topics/basics-of-iaea-safeguards; *More on Safeguards agreements*, IAEA, https://www.iaea.org/topics/safeguards-legal-framework/more-on-safeguards-agreements.

[333] For a focus on the latter, see Part II.

[334] Shavit, *Compute Monitoring*, at 4–5.

[335] *See* IAEA Statute art. VII.C. (establishing a similar staff for the IAEA).

[336] *Compare* NPT art. III.1 *with* NPT art. III.2; *see also* Laura Rockwood, *Legal Framework for IAEA Safeguards*, IAEA, at 4–5, https://www.iaea.org/sites/default/files/16/12/legalframeworkforsafeguards.pdf.

[337] James M. Acton, *Governance of Dual-Use Technologies: Theory and Practice, Ch. 1: On the Regulation of Dual-Use Nuclear Technology*,

regulation for AI-related materials–possibly, the U.S. and its allies and non-allied states, such as China and Russia–further research is needed on methods for ensuring safety for all of humanity and whether that would entail having a two-tiered system.

3. Complaints

For complaint mechanisms, despite the treaty's lack of institutional support entities, the Biological Weapons Convention (BWC) offers a potential model, permitting complaints for grievances attributable to a state Party to be lodged with the United Nations Security Council (UNSC).[338] Notably, the BWC has been critiqued for its failure to prevent the development of prominent biological weapons programs, such as that of the former Soviet Union.[339] Further, the BWC failed to create institutional entities meant to advance its objectives, not even possessing a Secretariat-like body until the Implementation Support Unit was created in 2006, a gap of over 30 years from the year of the treaty's widespread ratification.[340] While valid, critiques of the BWC should not merely focus on its complaint mechanism to the UNSC; rather, critiques should consider the treaty's lack of institutional entities and also that such a complaint mechanism may be useful in another context if supplemented with enforcement mechanisms.

Thus, a similar complaint mechanism for high-risk AI could permit certain violations to be reported to the UNSC or CHAI. For complaints to the UNSC, there are possible complaints that would be in every country's interest to address–such as those relating to misuse of powerful AI models with the capabilities to harm persons globally. Similarly, given the possible technical nature of some complaints, CHAI or the GOHAI Secretariat may be needed to provide explanations for certain risks or even to filter prominent complaints that should draw the attention of the UNSC.

4. Enforcement mechanisms for private entities

With many private entities controlling or developing high-risk AI,[341] an AI-safety treaty should ensure that there are sufficient enforcement mechanisms against private entities that undermine or act against the goals behind the treaty's objectives. One possible way to apply such

---

https://www.amacad.org/publication/governance-dual-use-technologies-theory-and-practice/section/4 (citing arts. VI and IV of the NPT).

[338] Biological Weapons Convention art. VI; *see also* footnote 206.

[339] *See* Jonathan B. Tucker, *Biological Weapons in the Former Soviet Union: An Interview with Dr. Kenneth Alibek,* at 4, https://www.nonproliferation.org/wp-content/uploads/npr/alibek63.pdf.

[340] United Nations, Office for Disarmament Affairs, *Implementation Support Unit*, https://disarmament.unoda.org/biological-weapons/implementation-support-unit/; Biological Weapons Convention, NTI, https://www.nti.org/education-center/treaties-and-regimes/convention-prohibition-development-production-and-stockpiling-bacteriological-biological-and-toxin-weapons-btwc/.

[341] *See, e.g.*, Fabio Urbina et al., *Dual use of artificial-intelligence-powered drug discovery*, 4 Nature Machine Intelligence 189 (2022); *see* Kevin Roose, *A.I. Poses 'Risks of Extinction," Industry Leaders Warn*, N.Y. Times (May 30, 2023), https://www.nytimes.com/2023/05/30/technology/ai-threat-warning.html.

accountability to private entities is through requiring state Parties to enact legislation or measures to hold private entities responsible through domestic civil or criminal law.[342] Many states incorporate a dualist perspective of international law, meaning that domestic legislation is needed to give effect to obligations undertaken in a treaty; thus, enacting legislation for a treaty could incorporate civil or criminal penalties for private entities.[343] State Parties that do not enact such civil or criminal penalties could further be held responsible under enforcement mechanisms applicable to state Parties.[344]

The Drug Treaties provide possible precedents of ensuring state Parties incorporate in their domestic criminal law penalties for private persons whose actions undermine or go against the purpose of the treaties. In particular, each of the Drug Treaties requires state Parties to establish as criminal offenses situations where a person acts "intentionally" while committing any of a list of enumerated actions.[345] The SCND specifies actions, among others, including "cultivation, production, manufacture, extraction . . . importation and exportation of drugs contrary to the provisions of the Convention . . ."[346]

Further, the Drug Treaties include provisions meant to ensure that state Parties are able to effectively penalize violations of domestic laws and regulations that enact the provisions of the treaty, as well as to further the purposes of the treaty. For instance, CAIT includes provisions ensuring the establishment of jurisdiction, of extradition when needed, and of mutual legal assistance among state Parties.[347] CAIT further requires state Parties to adopt measures enabling the confiscation of proceeds from criminal offenses enumerated in article 3, paragraph 1, as well as for the confiscation of the regulated substances themselves.[348]

Applying measures from the Drug Treaties to high-risk AI global governance should be approached with caution, particularly in considering which actions may be criminalized. However, with this significant caveat, treaty drafters could consider provisions that would permit carefully chosen criminal penalties to be effectively enforced. For example, to the extent that such actions are not already criminalized in state Parties' domestic legislation, actions taken to design biological weapons with high-risk AI models could be widely prohibited absent a license requiring high safety standards and personnel pre-clearance. Also, based loosely on CAIT, a high-risk AI safety treaty could include similar provisions meant to enable effective penalization of violations of criminal law, as well as permitting confiscation of proceeds from criminal use of certain AI models, for example.[349]

---

[342] Another approach outside the scope of this Article would be to explore methods of providing for private entity-state arbitration to resolve disputes relating to high-risk AI.
[343] The United Kingdom, for instance, possesses a dualist approach to international law, requiring an act of Parliament or judge-made law in the common law for international obligations to apply under domestic law. Lord Jonathan Hugh Mance, *International Law in the UK Supreme Court*, Feb. 13, 2017, https://www.supremecourt.uk/docs/speech-170213.pdf.
[344] See Part II.E.1.
[345] SCND art. 36; VCPS art. 22; CAIT art. 3.
[346] SCND art. 36(1).
[347] CAIT arts. 4, 6, 7.
[348] CAIT art. 5.
[349] *See* footnotes 345 and 346.

Similarly, the Cartagena Protocol and the Nagoya-Kuala Lumpur Supplementary Protocol on Liability and Redress to the Cartagena Protocol (together, the "Biological Diversity Protocols") provide possible guidance for implementation of both criminal and civil liability into state Parties' domestic law. The Cartagena Protocol requires state Parties to adopt civil measures and to consider criminal measures to prevent or penalize "transboundary movements of living modified organisms [LMOs] carried out in contravention of . . . domestic measures to implement this Protocol."[350]

The Nagoya-Kuala Lumpur Supplementary Protocol offers further tools for effectiveness, requiring any entity in control of a LMO to inform its respective national authority of any damage caused by that LMO; both the entity and national authority must then evaluate the damage and take appropriate responsive measures.[351]

Provisions similar to those of the Biological Diversity Protocols may also be useful for high-risk AI global governance. At a minimum, private entities in control of high-risk AIs should be obligated to report damages from models under their control to national authorities, which should be obligated to consider responsive mitigation measures.[352] However, common situations may occur where it is difficult to attribute causality to harms resulting from a tenuous connection with a model, and further research is likely needed to consider liability regimes for such situations.[353]

Finally, in the context of trade, CITES provides a model through which state Parties agreed to criminalize certain trade of the subject of regulation: for CITES, these subjects are endangered species.[354] Notably, given the success of CITES toward its objectives,[355] its components may offer promising precedents for global AI governance.

For global governance of high-risk AI, an underinclusive approach could request state Parties to impose criminal or civil liability for intentional trade of items that are in contravention of the treaty, such as certain ML chips, lithography equipment, and other components used in manufacturing ML chips.[356] As explored in Part III.B.2, this approach may be promising yet still underinclusive, as many high-risk models may result from fine-tuning or model prompting of existing models.

---

[350] Cartagena Protocol art. 25.
[351] Nagoya-Kuala Lumpur Supplementary Protocol on Liability and Redress to the Cartagena Protocol art. 5.
[352] *Compare with* Nagoya-Kuala Lumpur Supplementary Protocol on Liability and Redress to the Cartagena Protocol art. 5.
[353] For instance, an AI lab may release the model weights and all source code for a highly capable foundation model that is then subsequently fine-tuned by an unknown actor to cause immense harm.
[354] CITES art. VIII(1)–(2) (requiring state Parties to criminalize prohibited trade in violation of the Convention and providing the option of creating internal reimbursement methods for expenses incurred in confiscating specimens subject to such illegal trade).
[355] *CRS Report*, at 11 ("no species listed under CITES within the last 30 years has gone extinct.").
[356] *See* Gregory C. Allen, *Choking off China's Access to the Future of AI* (Oct. 2022), Center for Strategic & International Studies, https://csis-website-prod.s3.amazonaws.com/s3fs-public/2023-04/221011_Allen_China_AccesstoAI.pdf?VersionId=V2RCmPjpR8nQOybvB2LmNIu1Yx6RlYvA.

Conclusion

This Article has argued that high-risk AI models pose existential threats to humanity and that global governance is necessary to mitigate these risks. It has further made several unique contributions to the existing literature and to global AI governance efforts.

First, based on Jonas Schuett's factors, this Article has argued that a risk-based scope to global AI regulation is best, noting prominent examples of different scoping approaches using other methods. It has also argued for a framework of establishing global governance regimes based on (i) framing risks, (ii) determining objectives intended to address such risks, and (iii) enacting mechanisms under legal regimes capable of fulfilling those objectives. This approach is useful for both global and domestic regulatory regimes.

Second, after examining existing approaches to future possible harms, including risk, systemic risk, varieties of the precautionary principle, and uncertainty frameworks, this Article concludes that regulators in all scenarios must assign probabilities to the likelihood of future harmful events. With high-risk AI, regulators must eventually make such probability estimates before enacting informed regulatory measures.

Third, this Article has identified ideal characteristics of legal regimes for global AI safety measures, as well as existing regimes that share such characteristics. Further research may rely on these characteristics to identify applicable precedents for non-safety related objectives or to propose alternative AI safety regimes or measures.

Fourth, based on analyses of existing possible precedent global treaties and entities, this Article has suggested entities, organs, and functions of a global system meant to promote safety from high-risk AI.

Fifth and finally, based on precedent accountability mechanisms in international law, this Article explains possible transparency, oversight, complaints, and enforcement mechanisms for state Parties and private entities.

To meet the prominent challenges high-risk AI poses to humanity, further research is necessary. This Article did not explore the political challenges to constructing global governance regimes, nor did it focus on the best methods to ensure high-risk AI systems are secure from persons or entities with ill intent. Further research on methods of uniformly updating binding, global safety measures for high-risk AI is also needed, as well as optimal methods for amending a high-risk AI treaty.

It is my hope that this Article contributes to global efforts to promote prosperity and safety, while also permitting all conscious beings to experience the immense positive potential of AI. I welcome critiques of the frameworks and possible policy choices outlined in this Article, hoping that such a dialogue will lead to high-risk AI governance optimized for global well-being.