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 proposes objectives and ideal features of a global governance regime focused on unaligned AGI and other high-risk artificial intelligence (AI). It utilizes a risk-based approach to suggest mechanisms and legal frameworks that address two objectives: safety and accountability. Along with noting ideal characteristics of a global AI safety regime, this Article evaluates precedents in existing, comparable international systems to explore possible components of and entities for global governance of high-risk AI.

TABLE OF CONTENTS

[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 technical guidance 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 advancing 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 defined in various

[2] An "AI system" may be defined as "a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments . . . [.]" Artificial Intelligence [AI] Act, Regulation 2024/1689, 2024 O.J. (L 202) 1, art. 3(1) [hereinafter EU AI Act]. Regulation 2024/1689 of the European Parliament and of the Council of 13 June 2024 (Artificial Intelligence Act), 2024 O.J. (L 1689) 1, https://www.cambridge.org/core/journals/international-legal-materials/article/regulation-20241689-of-the-eur-parl-council-of-june-13-2024-eu-artificial-intelligence-act/64F1F6734F8C66CA3EEA149C9759194E [https://perma.cc/XZV2-MKPY].

[3] *See* Dylan Matthews, *The $1 Billion Gamble to Ensure AI Doesn't Destroy Humanity,* VOX (Sept. 25, 2023), https://www.vox.com/future-perfect/23794855/anthropic-ai-openai-claude-2 [https://perma.cc/4H6Z-T6KJ]; Google DeepMind, *About Google DeepMind*, Google DEEPMIND, https://www.deepmind.com/about/ (stating that their long-term aim is to solve intelligence, developing more general and capable problem-solving systems, known as artificial general intelligence (AGI)) [https://deepmind.google/about/]; Sam Altman,

ways, 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] Among different countries that may be in an AGI race, the United States and its allies' AI labs currently have a dominant position in AGI development.[5] The possible benefits of AGI are almost innumerable—among other offerings, AGI would likely help humanity fight negative effects of climate change and discover methods of preventing or mitigating harm from future pandemics.[6]

In pursuing AGI, many AI labs have already created highly capable, narrow AI that is exceptional in specific tasks and benefit 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 find new medicines, understand diseases, and more quickly 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 AI may democratize 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

---

Planning for AGI and Beyond, OPENAI (Feb. 24, 2023), https://openai.com/index/planning-for-agi-and-beyond [https://perma.cc/3LGX-KRNG] ("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)*, LINKEDIN, (July 7, 2023) https://www.linkedin.com/pulse/rise-artificial-general-intelligence-agi-xenonstack/ [https://perma.cc/AZ3S-XVXL].

[5] Haydn Belfield & Christian Ruhl, *Why Policy Makers Should Beware Claims of New "Arms Races,"* BULL. OF 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) [https://perma.cc/7YD7-2HBX]; *see also* Seán ÓhÉigeartaigh, *The Most Dangerous Fiction: The Rhetoric and Reality of the AI Race, in* THE ARTIFICIAL INTELLIGENCE REVOLUTION: CHALLENGES AND OPPORTUNITIES (Max Rangeley & Nicholas Fairfax, eds., 2026) , https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5278644 (noting that an AI race narrative may become a self-fulfilling prophecy that inhibits international cooperation on AI safety).

[6] STANFORD ONLINE, *Andrew Ng: Opportunities in AI – 2023*, at 35:36 (YouTube, Aug. 29, 2023), https://www.youtube.com/watch?v=5p248yoa3oE [https://perma.cc/8UMC-MC26].

[7] Google DeepMind, *AlphaFold,* GOOGLE DEEPMIND, https://www.deepmind.com/research/highlighted-research/alphafold [https://perma.cc/24B8-6L4V].

[8] *See id.*

risks to humanity. Many leading AI researchers 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 as dire as extinction.[10] Leading AI technical researchers Hendrycks, Mazeika, and Woodside further argue that AI can be a source of extreme risk, including: from malicious use, such as AI assistance in creating novel pathogens for bioterrorism; from overreliance on flawed AI in autonomous warfare, including the risk of generating potential nuclear conflicts; from organizational risks; and from rogue AI with actions that 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] International cooperation imposing uniform

---

[9] RICHARD NGO, AGI SAFETY FROM FIRST PRINCIPLES 1 (Sept. 2020); Dan Hendrycks & Mantas Mazeika, *X-Risk Analysis for AI Research*, ARXIV 1, 5 (2022), https://arxiv.org/pdf/2206.05862.pdf; Eliezer Yudkowsky, *Pausing AI Developments Isn't Enough. We Need to Shut It All Down*, TIME (Mar. 29, 2023), https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough/?ref=campaignforAIafety.org [https://perma.cc/EP8S-5Y3B].
[https://perma.cc/H25W-2X6H]; Benjamin S. Bucknall & Shiri Dori-Hacohen, *Current and Near-Term AI as a Potential Existential Risk Factor*, AIES 1,1 (Sept. 21, 2022), https://arxiv.org/pdf/2209.10604.pdf; DAN HENDRYCKS, MANTAS MAZEIKA, & THOMAS WOODSIDE, AN OVERVIEW OF CATASTROPHIC AI RISKS 2 (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/ [https://perma.cc/5Y5G-UEQ2]; AI SEOUL SUMMIT, INTERNATIONAL SCIENTIFIC REPORT ON THE SAFETY OF ADVANCED AI 83 (2024), https://assets.publishing.service.gov.uk/media/6716673b96def6d27a4c9b24/international_scientific_report_on_the_safety_of_advanced_ai_interim_report.pdf [https://perma.cc/53LF-NKA9].

[10] TOBY ORD, THE PRECIPICE: EXISTENTIAL RISK AND THE FUTURE OF HUMANITY 6 (2020) (In line with Ord's definition, Martínez and Winter found 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 number of lives to be existential); Eric Martínez & Christoph Winter, *Ordinary Meaning of Existential Risk* (Legal Priorities Project Working Paper, Paper No. 7-2022), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4304670.

[11] HENDRYCKS, MAZEIKA, & WOODSIDE, *supra* note 9, at 2, 5.

[12] *See* Yonadav Shavit, *What Does it Take to Catch a Chinchilla? Verifying Rules on Large-Scale Neural Network Training via Compute Monitoring*, ARXIV 1, 14 (2023), https://arxiv.org/pdf/2303.11341.pdf; Michelle Toh & Yoonjung Seo, *OpenAI CEO Calls for Global Cooperation to Regulate AI*, CNN BUSINESS (June 9, 2023), https://www.cnn.com/2023/06/09/tech/korea-altman-chatgpt-ai-regulation-intl-hnk/index.html [https://perma.cc/YL9R-NXRL].

protective safety measures would facilitate wider adoption of safety practices, as actors would know that 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, as preventing the arrival of such a rogue AGI would be in every state's interest.[14]

While there has been some movement toward international cooperation, further work is needed to mitigate AI risks. At a national level, both the United States and the United Kingdom have established respective centers focused on enabling governance of high-risk AI systems and on evaluating AI models for cyber, chemical, biological, and agent capabilities.[15] Further, while China has not stated it will advocate for global cooperation, it has recognized the existential threat of AI.[16]

At the international level, following two summits on AI safety and governance, countries at the forefront of AI signed non-binding declarations recognizing the potential for "catastrophic[] harm, either deliberate or unintentional," from AI.[17] Notably, while China, the United States, and the United Kingdom signed the Bletchley Declaration from the first summit, China neither signed the AI Seoul Summit's Seoul Declaration nor attended its leaders' session. Apart from these summits, the Global Partnership on Artificial Intelligence—with a Council, Steering Committee, and Secretariat hosted by the Organization for Economic Co-

---

[13] ALLAN DAFOE, AI GOVERNANCE: A RESEARCH AGENDA 46 (2018), https://cdn.governance.ai/GovAI-Research-Agenda.pdf.

[14] *See* TED, *Will Superintelligent AI End the World?*, at 4:51 (YouTube, July 11, 2023), https://www.youtube.com/watch?v=Yd0yQ9yxSYY [https://perma.cc/TPP4-XSMS] (arguing for international bans on certain AGI development).

[15] AI SEC. INST. [AII], https://www.aisi.gov.uk/ [https://perma.cc/C8YL-XAHV]; NAT'L INST. OF STANDARDS & TECH. [NIST], *Center for AI Standards and Innovation (CAISI)*, https://www.nist.gov/caisi [https://perma.cc/R42W-R5CA].

[16] Lauren Sforza, *China Warns of "Complicated and Challenging Circumstances" Posed by AI Risk,* HILL (May 31, 2023), https://thehill.com/policy/technology/4028141-china-warns-of-complicated-and-challenging-circumstances-posed-by-ai-risk/.

[17] GOV.UK, *The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023*, (Feb. 13, 2025), https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023 [https://perma.cc/UMA3-2XC7] (noting the need for international cooperation to address AI risks); *see also* SEOUL DECLARATION FOR SAFE, INNOVATIVE AND INCLUSIVE AI BY PARTICIPANTS ATTENDING THE LEADERS' SESSION OF THE AI SEOUL SUMMIT, 21ST MAY 2024 1 (2024), https://www.mofa.go.kr/viewer/skin/doc.html?fn=20240523095941078.pdf (same).

operation and Development, and two research centers—acts as a "multistakeholder expert community that brings together governments, industry, academia, and civil society . . . to advance human-centric and trustworthy AI."[18]

Nations have also signed onto binding international agreements. The most significant to date is likely the Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (the AI Convention), which has been signed but not yet ratified by the United States and most European countries, including the United Kingdom.[19] The AI Convention seeks to ensure that "activities within the lifecycle of [AI] systems are fully consistent with human rights, democracy and the rule of law," mandating the adoption of measures to protect those goals, including: transparency and oversight; accountability and responsibility; safe innovation; and procedural safeguards.[20] Based on a risk and impact management framework, the AI Convention requires parties to adopt measures that are "graduated and differentiated, as appropriate," listing certain requirements for such measures, including monitoring of risks, documentation of risks, and testing.[21] However, the AI Convention does not specify a certain level of monitoring or testing, and its vague wording thus leaves signatories free to implement light-touch measures that do not adequately account for catastrophic risks from AI.[22] Also, the AI Convention does not apply to matters relating to the protection of signatories' national security interests, to research and development of AI systems not yet deployed, or to matters relating to national defense.[23] In sum, current global governance of high-risk AI lacks effectiveness.

---

[18] ORG. FOR ECON. COOP. & DEV. [OECD], *About the Global Partnership on Artificial Intelligence (GPAI)*, https://oecd.ai/en/about/about-gpai [https://perma.cc/L4TQ-3VYB].

[19] *Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law*, COUNCIL OF EUROPE TREATY SERIES- NO. 225 (Sept. 5, 2024), https://rm.coe.int/1680afae3c.

[20] Many of these requirements are general mandates reliant on these end-goal objectives, not specific tailored requirements or minimum floors to ensure jurisdictional consistency in how those objectives are satisfied. *See id.* art. 1.1, arts. 8, 9, 13, 15.

[21] *Id.* art. 16(2).

[22] *See id.*

[23] *Id.* art. 3(2)–(4). While it may be difficult to have countries agree to a treaty concerning their national security or defense, the exception regarding research and development (R&D) is particularly concerning when accounting for the containment problem. Certain frontier AI systems may display behaviors during R&D, such as lying to their developers about the AI's safety or blackmailing an engineer when put at risk of being turned off, that pose risks to the

Further, scholarship on proposals for global governance regimes of high-risk AI is in progress but remains incomplete. While certain researchers and commentators have proposed general structures[24] or specific mechanisms[25] useful for possible AGI global governance organizations, many of these proposals either do not explain the functioning of such organizations or are not structured 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 warned of significant differences between the two technologies and of the need to tailor governance to the underlying technology.[26] More research is needed to explore ideal structures for possible functions and organs of a global governance body addressing AGI, considering a variety of important objectives.[27]

This Article proposes a new, tailored global governance system for certain existential risks from high-risk AI. It links a global AI governance regime's objectives to existential risks from AI, specifying possible entities and mechanisms to pursue objectives relating to high-risk AI. It recognizes but does not focus on the urgent need for dedicated global cooperation on AI safety research.[28] Further, this Article recognizes

---

very same objectives of the AI Convention. *See* ANTHROPIC, SYSTEM CARD: CLAUDE OPUS 4 & CLAUDE SONNET 4 28 (2025), https://www-cdn.anthropic.com/6d8a8055020700718b0c49369f60816ba2a7c285.pdf; *see also* James Babcock et al., *Guidelines for Artificial Intelligence Containment*, ARXIV 1, 5–6 (2016), https://arxiv.org/pdf/1707.08476.pdf (discussing guidelines to assist with sandboxing software development to safely evaluate AI).

[24] *See, e.g.*, Lewis Ho et al., *International Institutions for Advanced AI*, ARXIV 1, 1 (2023), https://arxiv.org/pdf/2307.04699.pdf.

[25] *See, e.g.,* Shavit, *supra* note 12, at 1, 14.

[26] *See, e.g.,* Ian J. Stewart, *Why the IAEA Model May Not Be Best for Regulating Artificial Intelligence*, BULLETIN OF THE ATOMIC SCIENTISTS (June 9, 2023), https://thebulletin.org/2023/06/why-the-iaea-model-may-not-be-best-for-regulating-artificial-intelligence/ [https://perma.cc/9YBK-J2G8].

[27] For a literature review of possible models of AI institutions, see Matthijs Maas & José Jaime Villalobos, *International AI Institutions: A Literature Review of Models, Examples, and Proposals*, SSRN Electron. J., Paper No. 4579773 (Sept. 2023), https://ssrn.com/abstract=4579773.

[28] For such proposals, see Ho et al., *supra* note 24, at 1; DANIEL ZHANG ET AL., ENHANCING INTERNATIONAL COOPERATION IN AI RESEARCH: THE CASE FOR A MULTILATERAL AI RESEARCH INSTITUTE, STANFORD UNIVERSITY HUMAN-CENTERED ARTIFICIAL INTELLIGENCE 5 (2022), 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*

but does not fully cover the additional possible need for global governance of non-existential risks from AI.[29] Rather, it provides 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: (1) maintaining and enforcing protective safety measures for high-risk AI; and (2) 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. It argues that a tiered, risk-based approach to global governance of high-risk AI is optimal since this approach would ensure that the regime is not needlessly intrusive to minor AI applications and is clear to its subjects of regulation.

Part II provides an overview of the potential existential risk from unaligned, highly capable AI 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 proposes 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.

Given the fast-moving nature of global AI initiatives, technological progress, and U.S. policy, certain recent events, policy choices, and initiatives may not be reflected herein.

Politically Neutral Hub for Basic AI Research, 7 CSS POL'Y PERSP. 1, 3 (2019), https://css.ethz.ch/content/dam/ethz/special-interest/gess/cis/center-for-securities-studies/pdfs/PP7-2_2019-E.pdf.

[29] For a non-exhaustive list of such risks, see Bernard Marr, *The 15 Biggest Risks of Artificial Intelligence*, FORBES (June 2, 2023), https://www.forbes.com/sites/bernardmarr/2023/06/02/the-15-biggest-risks-of-artificial-intelligence/?sh=7cf25eb82706 (last visited Apr. 14, 2026).

I. A BACKGROUND ON AI AND RISK-BASED SCOPES FOR GOVERNANCE

To provide context for possible points of regulation, this Part provides an overview of the process of developing powerful AI that may pose existential risks. It covers deep learning, 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 AI most likely to become AGIs,[30] 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 (DL), which currently drives AI development. DL refers to a subset of machine learning (ML)[31] through which neural networks, or processing nodes organized into deep layers, process large amounts of data to provide outputs based on identified probabilistic patterns.[32] 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.[33]

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.[34] 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

---

[30] *See* Sébastien Bubeck et al., *Sparks of Artificial General Intelligence: Early Experiments with GPT-4*, ARXIV 1, 4 (Apr. 13, 2023), https://arxiv.org/abs/2303.12712.

[31] Machine learning (ML) 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).

[32] Laurie A. Harris, Cong. Rsch. Serv., R46795, *Artificial Intelligence: Background, Selected Issues, and Policy Considerations* 4, 27 (2021).

[33] *See* Ashish Vaswani et al., *Attention Is All You Need*, ARXIV 1, 8 (2017), https://arxiv.org/abs/1706.03762.

[34] 3BLUE1BROWN, *But What is a Neural Network? | Chapter 1, Deep learning*, at 3:25 (YouTube, Oct. 5, 2017), https://www.youtube.com/watch?v=aircAruvnKk [https://perma.cc/GN9T-G3XP].

well as a "bias" that impacts the threshold needed to be surpassed before a node activates.[35]

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.[36] While making such adjustments may eventually produce desired outputs, the functioning—and, by extension, control—of neural networks are poorly understood in part because of the human inability to account for a given parameter's effect on a given output, an issue referred to as the "black box problem."[37]

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 "model that is trained on broad data . . . that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks . . . ."[38] For instance, the systems behind OpenAI's ChatGPT (e.g., GPT-5.2) are foundation models with billions of parameters that produce conversational text as the desired outputs.[39] Text-generating foundation models are also referred to as LLMs, but the term "foundation models" is more inclusive of powerful models, as many models are "multimodal (e.g., possess visual capabilities)."[40]

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 either (ii-a) reinforcement learning from

---

[35] *See id.* at 11:06. Together, the weights and biases of an AI system constitute its "parameters." *Id.* at 8:51.

[36] Aihwarya v. Srinivasan, *Stochastic Gradient Descent–Clearly Explained*, MEDIUM (Sept. 7, 2019), https://medium.com/data-science/stochastic-gradient-descent-clearly-explained-53d239905d31.

[37] *AI's Mysterious 'Black Box' Problem, Explained*, UNIV. OF MICH.-DEARBORN (Mar. 6, 2023), https://umdearborn.edu/news/AI-mysterious-black-box-problem-explained.

[38] Rishi Bommasani et al., *On the Opportunities and Risks of Foundation Models* 1, 3 (Ctr. for Rsch. on Found. Models, Stanford HAI, 2021), https://arxiv.org/abs/2108.07258.

[39] *Id.* at 98. As of August 2023, ChatGPT, a conversational AI application 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-app-in-the-history-of-web-applications/?sh=1e2b1efe678c (last visited Apr. 14, 2026).

[40] Markus Anderljung et al., *Frontier AI Regulation: Managing Emerging Risks to Public Safety* 7 n.8 (arXiv preprint, July 6, 2023) [hereinafter Anderljung et al., *Frontier AI Regulation*].

human feedback (RLHF) or (ii-b) a process known as constitutional AI or Reinforcement Learning from AI Feedback.[41] In each step, the model is fine-tuned from the previous step's resulting model, changing its parameters.[42]

First, in generative pretraining, the model undertakes unsupervised learning[43] 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 pretraining.[44] 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).[45] The overwhelming majority of companies producing high-quality ML chips are based in the United States or its allied states, including Google, Nvidia Corporation, and Advanced Micro Devices, Inc.[46] Taiwan Semiconductor Manufacturing Corporation dominates manufacturing of advanced ML chips,[47] and the Dutch company ASML Holdings has a near-monopoly on advanced lithography machines, which are needed to produce cutting-edge ML chips.[48]

---

[41] ARI SEFF, *How ChatGPT Is Trained*, at 0:55 (YouTube, Jan. 24, 2023), https://www.youtube.com/watch?v=VPRSBzXzavo [https://perma.cc/F7WD-VFNU]; Yuntao Bai et al., *Constitutional AI: Harmlessness from AI Feedback* (arXiv preprint, Dec. 15, 2022) https://arxiv.org/pdf/2212.08073. See *infra* for explanations of these processes.

[42] SEFF, *supra* note 41, at 1:51.

[43] IBM, *What Is Unsupervised Learning?*, https://www.ibm.com/think/topics/unsupervised-learning [https://perma.cc/8KM9-6DNQ]. Unsupervised learning uses ML algorithms to analyze and cluster datasets, finding patterns in data without human intervention, while supervised learning uses labeled data.

[44] SEFF, *supra* note 41, at 6:57.

[45] Shavit, *supra* note 12, at 1.

[46] *See id.* at 13; *see also* SAIF M. KHAN & ALEXANDER MANN, GEORGETOWN CTR. FOR SEC. & EMERGING TECH., AI CHIPS: WHAT THEY ARE AND WHY THEY MATTER: AN AI CHIPS REFERENCE 3 (2020) (noting that developing "the same AI application using older AI chips or general-purpose chips can cost tens to thousands of times more").

[47] *Taiwan's Dominance of the Chip Industry Makes It More Important*, ECONOMIST (Mar. 6, 2023), https://www.economist.com/special-report/2023/03/06/taiwans-dominance-of-the-chip-industry-makes-it-more-important (last visited Apr. 15, 2026).

[48] Xinmei Shen, *Chinese Imports of ASML Lithography Chip-Making Machines Have Surged Past the Dutch Company's 2023 Estimates*, S. CHINA MORNING POST (Aug. 28, 2023), https://www.scmp.com/tech/tech-war/article/3232401/chinese-imports-asml-lithography-chip-making-machines-have-surged-past-dutch-companys-2023-estimates (last visited Apr. 15, 2026).

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.[49] 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 data center infrastructure.[50] Further, advancements in ML chips, processors, systems, and algorithms facilitate more efficient development of foundation models, reducing the time, compute, and capital needed to produce a given foundation model.[51] Thus, given these technological advancements, generative pretraining for a foundation model may become cheaper and more accessible as time progresses.

Second, after generative pretraining results in a raw base foundation model, fine-tuning is used to train the model to produce text that mimics human text, as well as to reduce the instances where the model provides harmful or offensive information.[52] In the first step of fine-tuning, supervised fine-tuning is applied to the model. In this process, the model is trained on 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.[53]

In the second step of fine-tuning, either RLHF or constitutional AI may be applied.[54] 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."[55] Based on

---

[49] "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." Dario Amodei, *Written Testimony of Dario Amodei, Ph.D. Co-Founder and CEO, Anthropic - For a hearing on "Oversight of A.I.: Principles for Regulation,"* 1, 2 (2023) [hereinafter Amodei Testimony], https://www.judiciary.senate.gov/imo/media/doc/2023-07-26_-_testimony_-_amodei.pdf.

[50] Zachary Champion, *Optimization Could Cut the Carbon Footprint of AI Training by Up to 75%*, UNIV. MICH. (Apr. 17, 2023), https://news.umich.edu/optimization-could-cut-the-carbon-footprint-of-ai-training-by-up-to-75/.

[51] Amodei Testimony, *supra* note 49, at 2.

[52] SEFF, *supra* note 41, at 0:55.

[53] Stuart Russell, *Written Testimony of Stuart Russell*, 1, 6 (2023) [hereinafter Russell Testimony], https://www.judiciary.senate.gov/imo/media/doc/2023-07-26_-_testimony_-_russell.pdf.

[54] *Id.*

[55] *Id.*

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.[56] While providing an effective manner of fine-tuning, RLHF is imperfect and has significant problems, including its reliance on manipulable, fallible human evaluators who may push the raw model in dangerous directions before resulting in a "safer" model.[57]

Instead of using RLHF, an AI lab may 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 . . . . [B]ut there is no guarantee that the machine-generated rankings are comparable to human feedback."[58] In constitutional AI, the AI replaces a human in evaluating a separate model's outputs based on the set of principles or criteria.[59]

After it is fine-tuned, an AI may reach the public in different manners. Commonly, an AI model is offered to the public or customers through a chat user interface or application programming interface, which is essentially an intermediary between two applications.[60] Outputs from an AI may also be subject to safety filters to mitigate errors or reduce risks from an AI. Alternatively, the developer of an AI may open source the model, releasing its parameters and relevant code, enabling others to replicate that model or further fine-tune it.[61] While many AI developers

---

[56] *Id.*

[57] Charbel-Raphaël, *Compendium of Problems with RLHF*, LESS WRONG (Jan. 29, 2023), https://www.lesswrong.com/posts/d6DvuCKH5bSoT62DB/compendium-of-problems-with-rlhf [https://perma.cc/97SD-VHYS] (explaining many potential issues with reinforcement learning from human feedback (RLHF), including the difficulty of rewarding the right objective without having the model attempt to maximize its reward in a potentially harmful way).

[58] Russell Testimony, *supra* note 53, at 6 n.14.

[59] Bai et al., *supra* note 41, at 1.

[60] Rem Darbinyan, *What Are AI APIs, and How Do They Work?* (Apr. 13, 2022), https://www.dataversity.net/what-are-ai-apis-and-how-do-they-work/ [https://perma.cc/96Y4-KXHH]. For example, the underlying AI model behind ChatGPT is accessible to users through an application programming interface (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.

[61] IGUAZIO, *What Are Open Source Models in Machine Learning?*, https://www.iguazio.com/glossary/open-source-model/#:~:text=Open%20source%20has%20always%20been,of%20its%20open%2Dsource%20license [https://perma.cc/P8KV-Z29N]. For further information on distributions of

have become increasingly hesitant to open source their models, Meta appears committed to open sourcing its foundation models.[62]

B. A Risk-Based Scope for Global AI Governance

This Part 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 Part 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 the scope, or which AI 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*."[63] Similarly, this Article's proposed global governance regime should rely on a single term to determine its regulatory scope only if there is a good legal definition for AI that pose existential risks.

Based on the legal traditions of the United States and the European Union, Schuett identifies requirements for legal definitions, arguing such requirements should be used to determine if a term-based approach is ideal for setting a governance regime's scope.[64] Schuett specifies six requirements, stating that legal definitions: (1) must not be overinclusive, which would include cases that are not in need of regulation based on the regulation's objective; (2) must not be underinclusive; (3) must be precise, enabling a determination of whether a "particular case falls under the definition"; (4) must be understandable, ideally based on a word's ordinary meaning; (5) should be practical, including clear legal elements; and (6) should be flexible, accommodating technical progress.[65]

Thus, it may be difficult to draft a legal definition for AI, especially for those that may pose existential risks. For example, the EU AI

high-capability AI, see Toby Shevlane, *The Artefacts of Intelligence: Governing Scientists' Contribution to AI Proliferation* (2023) (D.Phil thesis, University of Oxford), https://cdn.governance.ai/Shevlane,_Artefacts_of_Intelligence.pdf.

[62] *See* Melissa Heikkilä, *Meta's Latest AI Model Is Free for All*, MIT TECH. REV. (July 18, 2023), https://www.technologyreview.com/2023/07/18/1076479/metas-latest-ai-model-is-free-for-all/ [https://perma.cc/9BKM-P575]; META AI, *Models and Libraries*, https://ai.meta.com/resources/models-and-libraries/ [https://perma.cc/D9A8-PYD6].

[63] Jonas Schuett, *Defining the Scope of AI Regulations*, 15 L. INNOVATION & TECH. 1, 2 (2023) (emphasis added).

[64] *Id.* at 5–6.

[65] *Id.* at 5.

Act draft from November 2022 offered a definition for an "artificial intelligence system" that reflected 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.[66] 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 final EU AI Act does define "general purpose AI model":

> [A]n AI model, including where such an AI model is trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications, except AI models that are used for research, development or prototyping activities before they are placed on the market[.][67]

After using Schuett's requirements to evaluate the above definition's effectiveness for capturing AI capable of posing existential risks—or even solely AI capable of becoming AGI—it would be difficult to establish the scope of regulation based on this term. A key problem

---

[66] *Compare 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*, art. 3(1) (Nov. 3, 2022), https://artificialintelligenceact.eu/wp-content/uploads/2022/11/AIA-CZ-Draft-General-Approach-11-Nov-22.pdf (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."), *with 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*, art. 3(1) (Apr. 21, 2021) [hereinafter EU AI Act, Apr. 2021 Draft], https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF ("[S]oftware . . . developed with one or more of the techniques and approaches listed in Annex I . . . .").

[67] EU AI Act, *supra* note 2, art. 3(63).

would be overinclusiveness. The above definition would include past foundation models, such as GPT-3, which does not pose an existential threat given its limited capabilities.[68] Given finite resources of regulatory bodies and possible resistance from entities whose models are needlessly regulated, overinclusiveness weighs heavily against using the above term to define the scope of regulation.[69] It is true that the definition is largely understandable and flexible; however, it would be difficult to base a global governance regime on a definition that is both overinclusive and underinclusive.

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 their] objectives."[70] 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')."[71] He also notes that definitions can better meet his stated legal requirements through the use of exemptions to reduce overinclusiveness, catch-all definitions for future flexibility, and sunset clauses or built-in revision schedules to enable updates to scope definitions.[72] To better meet the requirements for legal definitions when determining the scope of regulation for certain definitions, regulators may use a combination of technical approaches, applications, capabilities, and other aforementioned tools, instead of relying on a singular term's definition.[73]

---

[68] *See* Nick Barney et al., *What Is GPT-3? Everything You Need to Know*, TechTarget (Jan. 27, 2025), https://www.techtarget.com/searchenterpriseai/definition/GPT-3 [https://perma.cc/XR7F-PUX8].

[69] *See generally* Fabio Urbina et al., *Dual Use of Artificial-Intelligence-Powered Drug Discovery*, 4 Nature Mach. Intelligence 189, 189 (2022) (describing an AI model that researchers could use to design novel chemical toxins).

[70] Robert Baldwin & Julia Black, *Driving Priorities in Risk-based Regulation: What's the Problem?*, 43 J.L. & Soc'y 565, 567 (2016).

[71] Schuett, *supra* note 63, at 15.

[72] *Id.* at 21, 25.

[73] *Id.* at 24–25.

Notably, despite revisions leading to more precise definitions, the EU AI Act largely relies on a risk-based approach as the foundation of its governance system, estimating risks based on threats to "public interests, such as health and safety and the protection of fundamental rights, including democracy, the rule of law and environmental protection as recognised and protected by Union law."[74] Based on these risks, the EU AI Act divides AI into tiers: those 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.[75] To determine which AI are prohibited or deemed high-risk, the EU AI Act designates certain categories of AI; for instance, high-risk AI include systems subject to any of twelve product safety regulatory regimes or falling under any of eight categories.[76]

However, as with all tiered, risk-based regimes, how much regulatory scrutiny applies to a system depends on the classification system itself—a potential disadvantage for flexibility given future technological advancements. While the EU AI Act does not currently facilitate additions to listed prohibited AI—apart from amendments to the Act itself—it empowers the European Commission to amend the list of high-risk AI as long as two conditions are fulfilled: (1) the 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.[77] Thus, this system may not be quick to incorporate prohibitions of extreme high-risk AI or quick to add high-risk AI that do not satisfy the two aforementioned conditions. Importantly, a risk-based global governance regime should ensure that its classification of risk tiers is amendable and does not establish needlessly sharp cliffs between tiers.

---

[74] Artificial Intelligence Act, *supra* note 2, at Preamble ¶ 8; *see id.* art. 5 (Prohibited AI Practices); *id.* at ch. III (High-Risk AI Systems); *id.* at ch. IV (Transparency Obligations for Providers and Deployers of Certain AI Systems).

[75] *See id.* at ch. II (Prohibited AI Practices); *id.* at ch. III (High-Risk AI Systems); *id.* at ch. IV (Transparency Obligations for Providers and Deployers of Certain AI Systems); *id.* at ch. V (General-Purpose AI Models).

[76] *Id.* at Annex III, art. 7(2)(j). For a concise summary of a near-final draft of the EU AI Act's tiers, scope, requirements, and sanctions, see CHARLOTTE SIEGMANN & MARKUS ANDERLJUNG, CENTRE FOR THE GOVERNANCE OF AI, THE BRUSSELS EFFECT AND ARTIFICIAL INTELLIGENCE: HOW EU REGULATION WILL IMPACT THE GLOBAL AI MARKET 3–5 (2022)
[hereinafter SIEGMANN & ANDERLJUNG, BRUSSELS EFFECT AND AI].

[77] Artificial Intelligence Act, *supra* note 2, art. 36.

In a potentially useful comparison, some international treaties regulating dual-use substances have also taken a risk-based approach, such as the Montreal Protocol on Substances that Deplete the Ozone Layer (Montreal Protocol), which regulates substances that belong to a similar grouping but have vastly different potentials for harm.[78] To set its scope, the Montreal Protocol defines "controlled substances" as those listed in its annexes, including 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."[79]

Thus, this definition reflects the risk-based, enumerated approach, including exemptions to reduce overinclusiveness. For its enumerated substances, the Montreal Protocol offers listed categories and varying requirements in its annexes with different exemptions, as well as listed products containing such substances.[80] 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.[81]

In another global regime for dual-use substances, the Vienna Convention on Psychotropic Substances (VCPS) uses a risk-based approach to regulate substances that belong to a similar grouping but have vastly different potential for harm.[82] Just as some AI offer both beneficial and harmful applications, some psychotropic substances have both beneficial and harmful applications.[83] Like a prior draft definition for AI

---

[78] Montreal Protocol on Substances that Deplete the Ozone Layer art. 2, Sept. 16, 1987, 1522 U.N.T.S. 29 [hereinafter Montreal Protocol]. For a definition of "dual-use substances," see Elisa D. Harris et al., *Governance of Dual-Use Technologies: Theory and Practice*, AM. ACAD. ARTS & SCIS [hereinafter Harris et al., *Governance of Dual-Use Technologies*], https://www.amacad.org/publication/governance-dual-use-technologies-theory-and-practice/section/3 [https://perma.cc/YD9A-5UZM] (providing multiple definitions for "dual-use substances," including the definitions adopted by the European Commission and United States).

[79] Montreal Protocol, *supra* note 78, art. 1(4).

[80] *Id.* at Annex A, B, C, D, E, & F.

[81] *Id.* arts. 2(9), 2(10), 6.

[82] Vienna Convention on Psychotropic Substances [hereinafter VCPS] art. 20, Feb. 21, 1971, 1019 U.N.T.S. 175.

[83] *See The Benefits and Risks of Taking Psychiatric Medications*, NEXUS OF HOPE, https://nexusofhope.com/the-benefits-and-risks-of-taking-psychiatric-

in the EU AI Act,[84] 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."[85] Regulatory requirements apply to substances based on their risk tiers, with Schedule I being the most restrictive classification.[86] Regulated substances are not determined by criteria but are rather specifically enumerated.[87] If needed, the VCPS permits the Commission on Narcotic Drugs of the Economic and Social Council of the United Nations (UN) to add a substance, following a determinative, scientific finding from the World Health Organization (WHO).[88]

In a different approach that incorporates a risk assessment process in one of its core governance mechanisms, the Cartagena Protocol on Biosafety (Cartagena Protocol) relies on a term's definition to determine its scope. This approach results in overinclusiveness of regulated substances.[89] In other words, the Cartagena Protocol demonstrates the possible problems that may arise in relying on a definition to determine the scope of regulation. To determine its regulatory scope, the Cartagena Protocol defines the terms "living modified organism"[90] and states its measures will "apply to the transboundary movement, transit, handling and use of all living modified organisms [LMOs] that may have adverse effects on the conservation and sustainable use of biological diversity, taking also into account risks to human health."[91] 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.[92]

---

medications/#:~:text=They%20also%20affect%20your%20feelings,serotonin%20%E2%80%93%20your%20happy%20hormone [https://perma.cc/JEZ8-W6HH].

[84] EU AI Act, Apr. 2021 Draft, *supra* note 66, at Title I.

[85] VCPS, *supra* note 82, art. 1(e).

[86] *See id.* art. 7.

[87] *See id.* art. 8.

[88] *See id.* art. 2(4)(b).

[89] *See generally* Cartagena Protocol on Biosafety to the Convention on Biological Diversity, Jan. 29, 2000, 2226 U.N.T.S. 208 [hereinafter Cartagena Protocol] (containing expansive definitions and scopes, resulting in over-inclusion of substances in its regulations).

[90] *Id.* art. 3(g) (defining "living modified organism" (LMO) as "any living organism that possesses a novel combination of genetic material obtained through the use of modern biotechnology").

[91] *Id.* art. 4.

[92] *See id.* art. 5 (excludes, notably, certain LMOs used in pharmaceuticals, thus making use of an exemption to reduce over-inclusiveness).

Even with an exemption for pharmaceuticals, the Cartagena Protocol's scope of regulation is overinclusive, requiring individual determinations of which LMOs pose possible "adverse effects."[93] To facilitate this determination, the Cartagena Protocol heavily relies on assessments of risk to inform decisions and provide states guidance on objectives, use of such assessments, general principles, methodology, and factors to consider when conducting risk assessments.[94]

To summarize a non-exhaustive list, Table 1 lists approaches that 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 (Basel Convention), art. 1.

Thus, given the widespread acceptance of risk-based approaches to regulation in international treaties, a risk-based approach to global AI governance would not be unusual.[95] Framing risk tiers for AI posing existential risks will likely require flexibility for updating covered AI models.[96] It may also require a mixture of Schuett's described technical approaches,[97] applications, and capabilities to determine the scope of

[93] *Id.* art. 1.
[94] *Id.* at Annex III.
[95] *See also* Basel Convention on the Control of Transboundary Movements of Hazardous Wastes and their Disposal [hereinafter Basel Convention], Mar. 22, 1989, 1673 U.N.T.S. 126 (using a risk-based approach); Bern Convention on the Conservation of European Wildlife and Natural Habitats [hereinafter Bern Convention], Sept. 19, 1979, 1982 U.N.T.S. 197 (same).
[96] *See* Anderljung et al., *Frontier AI Regulation*, *supra* note 40, at 34 (describing the incredible growth in ML chips' speed and algorithmic efficiency for model training).
[97] *See* Schuett, *supra* note 63, at 15–16. RLHF and constitutional AI are currently dominant technical approaches in fine-tuning AI that may become AGIs, but new manners of fine-tuning may emerge. Targeting foundation models built with these methods may thus become underinclusive over time. *See supra* Part I.A.

regulation.[98] For example, the tier for prohibited use of AI models could include AI with the capability to launch certain nuclear weapons without human input. The tier for high-risk AI could include narrow AI models used in pharmaceutical research that can design biological weapons,[99] as well as designated models that have their capabilities estimated based on the ML chips, algorithms, and training run times that contribute to their training.[100]

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 risk-based system, where a risk is framed based on the threat it poses to an objective.[101] Baldwin and Black acknowledge that this method of "moving from a statement of statutory objectives to a set of key risks . . . is not a mechanical process[, and] it involves a host of discretionary and value-laden decisions."[102] Instead of taking this approach, this Article assesses AI systems' well-known potential risks before determining objectives that address those risks by using well-studied risks and frameworks to relatively minimize discretion in deciding objectives. From these objectives, this Article then explores 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, or the difficulty of having an AI system do exactly what humans want it to do without any unintended consequences. It then outlines possible methods of approaching plausible future 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

[98] *See id.* at 17–19. Relying too heavily on capabilities to frame risk would be a mistake because humans are currently not good judges of an AI's capabilities. See *infra* Part II.A (describing deception and sudden, emergent capabilities from AI).
[99] *See infra*.
[100] For the latter group, a margin of safety and built-in adjustments may be needed, considering the difficulty in predicting models' capabilities and the rapid progression of frontier models.
[101] Baldwin & Black, *supra* note 70, at 573–74.
[102] *Id.* at 582.

then explores initial possible protective safety measures that would mitigate risks caused in part by unaligned AI.

Further, this Part analyzes eleven treaties to propose ideal characteristics of legal models for global governance of high-risk AI. Finally, based on existing legal models, this Part recommends possible components of a global safety system addressing high-risk AI.

A. The Alignment Problem

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 AI that will try to do what their operators want them to do.[103] Given that AI 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.[104] Further, even if outer alignment is solved, the inner alignment problem would need to be addressed, as AI might develop subgoals that differ from those conveyed by the objective function.[105] For instance, during training, certain subgoals, such as gathering resources, information, or power, may enable an AI to consistently attain high scores on an objective function, possibly resulting in an AI with a goal of acquiring power.[106]

In addition, inner alignment and outer alignment combined do not completely encompass the intent problem.[107] Even with a "'safe' objective function,"[108] the "'intention,' 'incentive,' or 'motive'" of an AI cannot be equated to the objective for which it is optimized.[109] Humans, for instance, are arguably optimized for the objective of reproductive fitness, yet many humans do not primarily act with intent to optimize

---

[103] Paul Christiano, *Clarifying "AI Alignment,"* MEDIUM (Apr. 7, 2018), https://ai-alignment.com/clarifying-ai-alignment-cec47cd69dd6 [https://perma.cc/KKZ4-UH4R].
[104] NGO, *supra* note 9, at 18.
[105] *Id.* at 19. In other words, outer alignment is "the problem of correctly evaluating AI behaviour; inner alignment is the problem of making the AI's goals match those evaluations." *Id.* at 21.
[106] *Id.* at 19.
[107] *Id.* at 21–23.
[108] *Id*. at 19.
[109] Christiano, *supra* note 103.

their reproductive fitness.[110] Along with problems arising from the lack of intent alignment, AI have demonstrated other problematic behaviors, such as emergent functionality of new capabilities or goals, as well as deception of human evaluators.[111] Given these unsolved and poorly understood risks from AI, the emergence of more powerful and capable AI that mimic existing problematic behavior of less capable AI would pose extreme risks.[112]

Despite these well-known dangers, AI labs in the United States are openly competing to develop AGI and are increasing amounts of funding toward this goal.[113] If the needed capabilities to train foundation models spread outside of the United States and lead to an AGI race among countries, these race dynamics may increasingly lead to a race to the bottom, characterized by decreasing attention to safety, 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 result in humanity becoming a "second species," or one that loses control of its future to unaligned AGIs.[114]

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 solve the alignment problem and other AI safety issues before an AGI is developed. At best, safety measures can stop certain dangerous developments altogether; for example, required security testing can lead to the discovery and removal of dangerous capabilities prior to deployment.[115]

B. Frameworks for Approaching Unaligned, High-Risk AI

To better understand possible approaches to safety measures under risk-based governance, this Part examines various methods of assessing unaligned, high-risk AI, including risk, systemic risk, the

---

[110] *Id*.
[111] Hendrycks & Mazeika, *supra* note 9, at 5.
[112] *Id.* at 1.
[113] *See supra* note 3.
[114] NGO, *supra* note 9, at 1.
[115] *See* FUTURE LIFE INST., *Pause Giant AI Experiments: An Open Letter* (Mar. 22, 2023), https://futureoflife.org/wp-content/uploads/2023/05/FLI_Pause-Giant-AI-Experiments_An-Open-Letter.pdf.

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 "[r]isk is defined as the possibility of loss or injury" and has roots in computational occupations, meaning it is "something to be measured, often mitigated, and taken into account."[116] She further comments that a formal risk analysis often involves calculations: risk is calculated by multiplying the likelihood of an event by the measurable harm to be caused by such an event.[117] In the AI existential risk literature, risk has been framed as the following: risk = (hazard severity and prevalence) x exposure x vulnerability.[118]

Risks often involve measured potential threats, so a risk's framing is crucial for subsequent risk-based regulatory regimes.[119] When approaching possible future harms through the risk framework, policymakers tend to be utilitarian, with a focus on: quantifiable harms; harms that result from muddled chains of causality and that are often externalities to those producing them; and possible harms arising out of actions undertaken with the potential for a beneficial outcome.[120] These common approaches may impose limitations, such as a default preference for more quantifiable harms over less quantifiable harms.[121]

Further, risk-based frameworks follow a set process for the construction of an oversight regime. Typically, the regulating entity does the following: (1) "sets the level and types of risk it will tolerate"; (2) conducts a risk assessment, including an assessment of the likelihood of the harm occurring; (3) "evaluate[s] the risk and rank[s] the regulated

---

[116] Margot E. Kaminski, *Regulating the Risks of AI*, 103 B.U. L. REV. 1347, 1390 (2023).
[117] *Id.* at 1392–93.
[118] 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, *supra* note 9, at 2.
[119] Baldwin & Black, *supra* note 70, at 579.
[120] Kaminski, *supra* note 116, at 1393.
[121] *Id.* at 1393; *see id.* at 1397 (noting harms that are either not quantifiable or difficult to quantify without arbitrary policy choices).

entities [or activities] on their level or tier of risk—high, medium or low"; and (4) allocates resources based on the tiers of risk.[122]

The structure of the EU AI Act reflects this process, as 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 AI models or specific AI systems;[123] (3) ranked those enumerated AI 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 AI.[124]

For existential risks from unaligned AI, applying this risk framework results in a few challenges. An existential risk involves a possible low-occurrence, high-consequence event that may be difficult to estimate with accuracy.[125] For such rare or unprecedented events, quantifying these risks poses difficulties for knowledge, measurements, and mitigation.[126]

Even with such limitations, the risk framework is a useful base for constructing global AI regulatory safety measures by addressing the alignment problem. First, the state 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, the Parties would also need to be careful to not needlessly

---

[122] Michael Guihot, Anne F. Matthew & Nicholas P. Suzor, *Nudging Robots: Innovative Solutions to Regulate Artificial Intelligence*, 20 VAND. J. ENT. & TECH. L. 385, 450 (2017) [hereinafter Guihot et al., *Nudging Robots*].

[123] "AI models" may refer to "the raw, mathematical essence that is often the 'engine' of AI applications," while "AI systems" may be "a combination of several components, including one or more AI models . . . designed to be particularly useful to humans in some way. For example, the ChatGPT app is an AI system; its core engine, GPT-4, is an AI model." U.K. DEP'T FOR SCI., INNOVATION & TECH. & AI SAFETY INST., INTERNATIONAL AI SAFETY REPORT 26–27 (2025).

[124] *See generally* Artificial Intelligence Act, *supra* note 2 (displaying the aforementioned steps in reasoning in setting their regulations); SIEGMANN & ANDERLJUNG, BRUSSELS EFFECT AND AI, *supra* note 76, at 15 (summarizing the main provisions of the Artificial Intelligence Act, including sanctions and requirements).

[125] Hendrycks & Mazeika, *supra* note 9, at 1 (noting that some current estimates for AI-posed existential threats are in the single-digits over the next century).

[126] Kaminski, *supra* note 116, at 1379. Despite these difficulties, many are beginning to attempt such estimates. *See, e.g.*, ORD, *supra* note 10, at 167 (estimating the likelihood of an existential catastrophe from unaligned AI within the next 100 years as 10%).

cabin regulation within a specific sector or AI application.[127] The framing of the types of risk to be addressed is subjective and heavily influenced by policy preferences,[128] 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 determining the likelihood and magnitude of existential harms to life or health from unaligned AI, the Parties may struggle to determine the "specific details" of how such a catastrophe may occur, while the high probability of such an accident may be easier to estimate.[129] Currently, any highly capable AI is guaranteed to be unaligned to an extent.[130] 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 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 AI, are useful proxies for estimating existential risks to life and health.[131] Assuming an AI will be unaligned, likelihood of risk would largely rely on estimates of an AI's projected capabilities.

Thus, the 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. Inputs like compute, ML chip quality, and training run time are good indicators of an AI's capabilities.[132] Initial input-based risk assessments would likely need to be calculated and planned by a multidisciplinary team of experts, resulting in risk tiers with built-in adjustments accounting for technological progress or progression of ML chips, algorithms, and other inputs, along with a built-in margin of safety.

---

[127] Alicia Solow-Niederman, *Administering Artificial Intelligence*, 93 S. CAL. L. REV. 633, 670 (2020) ("[C]abining [AI] regulation too firmly within particular sectors splinters imperative conversations about cross-cutting values and norms for all domains.").
[128] *See supra* Part I.B.
[129] *See* Solow-Niederman, *supra* note 127, at 663.
[130] *See* discussions *supra* Part I.A and Part II.A (describing the imperfect process of how to train a foundation model and the unsolved alignment problem).
[131] *See* Schuett, *supra* note 63, at 14–15.
[132] *See supra* Part I.A.

Third, the Parties would evaluate the risks and rank regulated entities or activities, creating tiers. Fourth, each of those tiers would ideally receive tailored, uniform global treatment, including possible bans for unaligned AI deemed too capable based on specific metrics, such as agentic abilities or biological risk indicators.

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.[133] Instead of framing risk as a singular entity, Baldwin and Black contend risk may be 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."[134] For instance, in studying existential risks, many no longer characterize such risks "as singular events . . . [, preferring] descriptions of risks as the result of the complex interactions between multiple, more mundane vulnerabilities in our social and political systems."[135] 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."[136]

Drawing on the above principles, systemic risks from unaligned, high-risk AI may be characterized at the level of the AI global ecosystem or at the level of the individual AI system. First, in assessing the global ecosystem for the development of powerful foundation models, regulators and researchers seeking to address risk may need to consider risk factors, such as state-state competition, state-corporation power relationships, and AI interactions with nuclear threats, among others.[137] 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

[133] *See* Guihot et al., *Nudging Robots*, *supra* note 122, 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.'" (alteration in original) (quoting Ortwin Renn & Andreas Klinke, *Systemic Risks: A New Challenge for Risk Management*, 5 EUR. MOLECULAR BIOLOGY ORG. REP. S41, S41 (2004))).
[134] Baldwin & Black, *supra* note 70, at 569.
[135] Bucknall & Dori-Hacohen, *supra* note 9, at 120–21.
[136] Hendrycks & Mazeika, *supra* note 9, at 3–4.
[137] *See* Bucknall & Dori-Hacohen, *supra* note 9, at 121, 124.

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: (1) an AI's internal characteristics, such as interpretability; (2) the external environment's effect on an AI, such as access to the internet; and (3) the AI's effect on the external environment, such as an AI influencing a nuclear weapon launch.[138] The complex system that is the AI model and its other components could make it difficult to determine the external environment's effect and "predict how a particular intervention will unfold . . . [or] specify a cause-and-effect relationship."[139] 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.[140]

Given a systemic view of existential risk from unaligned AI, 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, such as potential information hazards from safety evaluations of high-risk AI systems.

3. The Precautionary Principle

For situations when quantifying risks or arriving at probabilities is difficult, the precautionary principle may be invoked by global regulators of high-risk, unaligned AI. While a variety of manifestations exist, the precautionary principle can be broadly 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."[141] Generally, versions of the precautionary principle may be interpreted to: (1) not justify inaction (a minimal interpretation); (2) justify regulatory action even if the link between cause and effect is not completely established (median interpretation); or (3) necessitate regulation until the

---

[138] Schuett, *supra* note 63, at 14–15.

[139] Alicia Solow-Niederman, *supra* note 127, at 673–74.

[140] *See* Charbel-Raphaël, *supra* note 57 (using an example to show that optimizing a model for a given reward can lead the model to attempt to earn such a reward by any means necessary).

[141] Grant Wilson, *Minimizing Global Catastrophic and Existential Risks from Emerging Technologies Through International Law*, at 33 (Global Catastrophic Risk Inst.), https://gcrinstitute.org/papers/006_international-law.pdf.

absence of a hazard is proven.[142] Naturally, to affect an outcome, one must specify a version of the principle.

At its weakest, the precautionary principle states that "scientific uncertainty is an inappropriate reason *not* to take precautionary actions."[143] For example, the Cartagena Protocol cites Principle 15 of the Rio Declaration on Environment and Development, which states "[w]here 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."[144] 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.

Given the potential extinction- or catastrophic-level risks posed by unaligned AI, applying an aggressive, maximalist form of the precautionary principle to global regulation would 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."[145] 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.[146]

An even stronger framing of the precautionary principle useful for unaligned AI posing an extinction-level threat is Cass Sunstein's 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."[147] The cost of this "option is that of delaying the decision until better information is

---

[142] DIDIER BOURGUIGNON, THE PRECAUTIONARY PRINCIPLE: DEFINITIONS, APPLICATIONS AND GOVERNANCE 7 (2015), https://www.europarl.europa.eu/RegData/etudes/IDAN/2015/573876/EPRS_IDA(2015)573876_EN.pdf.

[143] Wilson, *supra* note 141, at 34.

[144] Rio Declaration on Environment and Development 1992 Principle 15, Apr. 25, 1997, 31 I.L.M. 874.

[145] Cass R. Sunstein, *Irreversible and Catastrophic*, 91 CORNELL L. REV. 841, 850 (2006) [hereinafter Sunstein, *Irreversible and Catastrophic*].

[146] *Id.*

[147] *Id.* at 845.

available."[148] Given the aforementioned lack of understanding about the possibility of an unaligned AGI, humanity could buy an option (for example, delay developments past certain capabilities thresholds) to protect against irreversible harm until the alignment problem is solved and prevent the development of AGI. This option's value would take into account the 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.

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 health, safety, and the environment. Second, supporting evidence overwhelmingly shows foundation models are unaligned to an extent.[149] Key uncertainty only remains as to whether AGI is possible, as well as to 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.

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.[150] Even if one argues that true uncertainty is very rare or nonexistent, "bounded uncertainty," a situation where one cannot assign probabilities within specific bands, is more likely to occur.[151] Thus, even under an uncertainty-based approach, policymakers may need to engage with the language of probability and risk.

---

[148] *Id.*

[149] *See, e.g.*, Roger Brent & T. Greg McKelvey, Jr., *Contemporary AI Foundation Models Increase Biological Weapons Risk* iv (RAND Working Paper No. WR-A3853-1, June 2025), https://arxiv.org/abs/2506.13798 (explaining how three large language models may meaningfully assist in the development of biological weapons, despite safety assessments and other guardrails implemented by developers).

[150] Sunstein, *Irreversible and Catastrophic*, *supra* note 145, at 848.

[151] Cass R. Sunstein, *The Catastrophic Harm Precautionary Principle*, 6 ISSUES IN LEGAL SCHOLARSHIP 1, 20 (2007).

In an approach for possible catastrophic situations with uncertainty, the concept of 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 severe consequence.[152] 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; thus, even accepting the worst consequence of banning AGI development, it is likely a better outcome than extinction.

However, choosing to apply maximin is not that simple. As when invoking the precautionary principle, applying maximin requires a minimum threshold of plausibility of the worst-case scenario occurring. Again, this approach requires the language of probability and risk.[153]

In the 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?"[154] 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.[155] 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.[156] 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 AI is implemented, policymakers would need to pursue other actions. While knowledge of AI may increase over time and enable better probability-based risk

---

[152] JON ELSTER, EXPLAINING TECHNICAL CHANGE: A CASE STUDY IN THE PHILOSOPHY OF SCIENCE
203 (1983).

[153] *See* Bucknall & Dori-Hacohen, *supra* note 9, at 6. A possible application of maximin would be to ban the creation of specified foundation models based on certain input ceilings, such as a ceiling of a certain level of compute.

[154] Sunstein, *Irreversible and Catastrophic*, *supra* note 145, 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, *supra* note 151, at 23–24.

[155] Sunstein, *Irreversible and Catastrophic*, *supra* note 145, at 879.

[156] *See* Sunstein, *supra* note 151, at 19.

calculations,[157] it is not clear that scientists working at the forefront of new technologies are good predictors of when technologies will manifest such prominent capabilities.[158] Further, an important component of addressing the alignment problem is ensuring that safety research results in a higher ratio of safety to capabilities, meaning that research contributes to understanding of and progression toward overall safety more than it moves AI closer to becoming AGIs (capabilities).[159] Thus, a potential ban on powerful AI could run into the issue of scientists not knowing how to comply with the ban for certain activities.

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 AI. 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."[160] The chance of the worst-case scenario from unaligned AI is estimated to be about ten percent in the next century, so it is reasonable to state it is not highly improbable.[161] Further, while most alternatives to extinction or mass catastrophe are clearly better outcomes than such mass suffering, it is not clear such alternatives are much more likely.

Finally, like the precautionary principle, Sunstein's proposed 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."[162] A second, stronger version of this principle would have regulators 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

---

[157] *Id.* at 20.

[158] *Cf.* Ashutosh Jogalekar, *Leo Szilárd, a Traffic Light and a Slice of Nuclear History*, SCI. AM. (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).

[159] Hendrycks & Mazeika, *supra* note 9, at 8–9.

[160] Sunstein, *Irreversible and Catastrophic*, *supra* note 145, at 879.

[161] ORD, *supra* note 10, at 167.

[162] Sunstein, *supra* note 151, at 3.

the deaths of millions or billions.[163] 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 addressing catastrophic risks.[164] Together, the approaches of uncertainty, risk, and the variations of the precautionary principle provide policymakers with various options for approaching global regulation of unaligned AI.

C. Protective Safety Measures for Unaligned, High-Risk AI

Building on the previous exploration of frameworks for assessing threats from unaligned, high-risk AI, this Part explores possible initial protective safety measures (mechanisms). These mechanisms may have procedural or substantive aspects.[165] Under a risk-tiered system, different safety measures could be established in a treaty's initial text for each tier.

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 AI. For high-risk AI systems, red teaming would involve hiring "external red teams to identify hazards in their AI systems to inform deployment decisions[,] . . . [such as exploring] dangerous behaviors or vulnerabilities in monitoring systems intended to prevent disallowed use."[166] Red teaming could also be used to "provide indirect evidence that an AI system might be unsafe."[167] In a survey of leading experts from AGI labs, academia, and civil society, 98% of respondents somewhat or strongly agreed that "AGI labs should commission external red teams before deploying powerful models."[168]

However, given the possible variety in backgrounds, skills, and resources of potential red teams, global minimal standards could facilitate

[163] *Id.* at 6.
[164] *Id.* at 7–8.
[165] For instance, the EU AI Act has substantive requirements for high-risk systems, such as data quality and accuracy, as well as procedural requirements, such as recordkeeping meant to facilitate transparency. Kaminski, *supra* note 116, at 1376, 1389.
[166] HENDRYCKS, MAZEIKA, & WOODSIDE, *supra* note 9, at 32.
[167] *Id.*
[168] Jonas Schuett et al., *Towards Best Practices in AGI Safety and Governance: A Survey of Expert Opinion*, ARXIV 1, 18 (2023), https://arxiv.org/pdf/2305.07153.pdf.

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: (1) sufficient experience in interacting with state-of-the-art AI models; (2) sufficient amount of access to the AI models being red teamed; (3) adequate resources for red teams; (4) communication of certain results with regulators as needed; and (5) including red teams in the process of post-deployment model updates.[169] Further, given that red teaming may be increasingly difficult for certain RLHF models as they scale in size,[170] cutting-edge methods may be necessary to ensure effectiveness, such as red teaming possibly unaligned AI through the use of aligned or less capable AI.[171]

Currently, METR (Model Evaluation and Threat Research, also formerly known as ARC Evals), a prominent red teaming organization, provides a preliminary model of possible methods for addressing threats posed by high-risk AI. METR 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.[172] METR is particularly concerned with misaligned AI's potential for autonomous replication, concentrating its current red teaming efforts on evaluating existing models' autonomous replication abilities.[173] In an evaluation of models based on LLMs from Anthropic (Claude) and OpenAI (GPT-4), METR found it unlikely that such models were capable of creating dangerous autonomous agents.[174] METR'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,

[169] Anderljung et al., *Frontier AI Regulation*, *supra* note 40, at 26.
[170] Deep Ganguli et al., *Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned*, ARXIV 1, 1 (2022), https://arxiv.org/pdf/2209.07858.pdf.
[171] Ethan Perez et al., *Red Teaming Language Models with Language Models*, ARXIV 1, 7 (2022), https://arxiv.org/pdf/2202.03286.pdf.
[172] METR, *About METR,* https://metr.org/about [https://perma.cc/LPM5-3FSV].
[173] *See* METR, *Details About METR's Evaluation of OpenAI GPT-5.1-Codex-Max* (Nov. 19, 2025), https://evaluations.metr.org/gpt-5-1-codex-max-report/ [https://perma.cc/Z5QF-6A4N].
[174] Beth Barnes, *ARC Evals New Report: Evaluating Language-Model Agents on Realistic Autonomous Tasks*, LESSWRONG (Aug. 1, 2023), https://www.lesswrong.com/posts/EPLk8QxETC5FEhoxK/arc-evals-new-report-evaluating-language-model-agents-on.

ranging from searching a filesystem for a password to creating a language model agent.[175]

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.[176]

However, these guiding standards likely need reinforcement through more stringent, binding standards applicable to high-risk AI development and deployment.[177] Researchers working with high-risk AI have 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.[178] In the global context, this reporting structure could be uniformly mandated by all states where high-risk AI models are being developed, requiring the reporting of certain activities to a respective national regulatory authority.

Third, given certain models' potential to demonstrate (or be modified to produce) new risks after deployment, policymakers could standardize globally uniform post-deployment obligations for providers of high-risk models. Depending on how a model is released or accessible to users, new dangerous capabilities can be further developed or unlocked through: (1) fine-tuning; (2) "chain-of-thought prompting," which involves telling a model to think through problems step-by-step to improve problem-solving capabilities; (3) enabling LLMs to use external tools; (4) "automated prompt engineering," which involves "[u]sing LLMs to generate and search over novel prompts that can . . . elicit better

---

[175] *Id.*

[176] MILES BRUNDAGE ET AL., THE MALICIOUS USE OF ARTIFICIAL INTELLIGENCE: FORECASTING, PREVENTION, AND MITIGATION 56 (2018), https://arxiv.org/pdf/1802.07228.pdf.

[177] *See* Matthew Hutson, *Who Should Stop Unethical A.I.?*, 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, *supra* note 9, at 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).

[178] Fabio Urbina et al., *supra* note 69, at 190.

performance on a task"; and (5) foundation model programs, which integrate "foundation models into complex programs."[179] Substantive post-deployment obligations could involve regular risk assessments, provision of incident reporting mechanisms, and mechanisms for quickly withdrawing a deployed model.[180]

2. Procedural Mechanisms

Along with substantive mechanisms, procedural mechanisms can help promote safety. For example, an ex ante mechanism of a globally mandated licensing process for high-risk models could help control high-risk AI development or deployment.[181] In the U.S. administrative law context, licensing involves an expert agency that may set safety standards for a technology or practice and then assesses the regulated entity prior to a public release.[182] This assessment may be relative to whether the technology, the practice, or an entity meets a performance standard, meaning this procedural mechanism incorporates a substantive threshold for safety.[183] 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.[184]

For global oversight of high-risk AI, the licensing process could require countries to license the deployment and development of high-risk AI.[185] 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.[186] 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.[187]

---

[179] Anderljung et al., *Frontier AI Regulation*, *supra* note 40, at 12.
[180] *Id.* at 28.
[181] *See* Ho et al., *supra* note 24, at 6.
[182] *See* 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).
[183] *Id.* at 107.
[184] Anderljung et al., *Frontier AI Regulation*, *supra* note 40, at 19.
[185] *Id.* at 20.
[186] *Id.*
[187] *Id.*

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.[188] Another desirable trait of a global licensing regime would be to establish conditional licensing, while mandating ongoing compliance with specified safety guardrails.[189] However, in establishing a licensing regime, global regulators would need to be cautious about entrenching centralized power with regulated companies and AI labs, as heightened regulation could increase costs for new market participants and startups.[190]

Another procedural mechanism—that is arguably a variation of a licensing system—is the requirement for a conformity assessment performed by a third party or regulated entity before a model is released, as well as after its release.[191] For example, the EU AI Act requires a successful conformity assessment from third-party private entities for designated high-risk AI models. Upon successful completion, these high-risk models may be designated as sufficiently safe through an EU "technical documentation assessment certificate."[192] The EU AI Act conformity assessment mandates compliance with requirements for a risk management system, recordkeeping, human oversight, and post-market monitoring, among others.[193] While a conformity assessment conducted by a non-governmental entity may be a useful complement to government licensing, private entities are not accountable to the public in the same manner as democratically elected governments.

Finally, the requirement for a prepublication risk assessment before the release of potentially dangerous information is a feasible procedural mechanism aimed toward safety, and it could be completed via a working paper or publication tied to a high-risk model. This prepublication risk assessment could mandate analysis of "particular

---

[188] *See* Gianclaudio Malgieri & Frank Pasquale, *From Transparency to Justification: Toward* Ex Ante *Accountability for AI* 14, Brooklyn Law School, Legal Studies Paper No. 712 (2022).
[189] Kaminski, *supra* note 116, at 1388.
[190] Anderljung et al., *Frontier AI Regulation*, *supra* note 40, at 31.
[191] Kaminski, *supra* note 116, at 1381–82 (describing the conformity process as "licensing 'ultra-lite,'" completed by third parties or the regulated entities themselves).
[192] EU AI Act, *supra* note 2, art. 45(1).
[193] *See generally* EU AI Act, *supra* note 2 (discussing throughout a regulatory regime that utilizes a compliance system requiring human oversight, recordkeeping, monitoring after market entry, and implementation of a risk management system). *See also* SIEGMANN & ANDERLJUNG, BRUSSELS EFFECT AND AI, *supra* note 76, at 15 (listing the aforementioned compliance requirements for high-risk systems).

risks . . . of a particular capability if it became widely available, and deciding on that basis whether, and to what extent, to publish it." [194] 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. [195]

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 Part seeks to identify relevant, desirable characteristics of such models. 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. [196] 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. [197] On the other hand, based on a technology's externalities, its dual-use nature can be framed as its potential for harmful purposes or human betterment. [198] For this characteristic, this Article uses 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—a regulatory regime that regulates specified similar

[194] BRUNDAGE ET AL., *supra* note 176,at 86.
[195] *See supra* Part I.A; HENDRYCKS, MAZEIKA, & WOODSIDE, *supra* note 9, at 32.
[196] "[T]he concept of dual-use technology is a useful organizing principle for examining governance efforts across different technology areas." Harris et al., *Governance of Dual-Use Technologies*, *supra* note 78, at 6.
[197] *Id.* at 4 (citing Council Regulation (EC) No. 428/2009 of May 5, 2009, O.J. (L134) 1, https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:134:0001:0269:en:PDF [https://perma.cc/Z9CT-CZ5Z]; 15 C.F.R. § 730.3 (2000)).
[198] *Id.* at 5.

substances or groups of technologies based on tiers of their respective risks[199]—is an effective approach for unaligned AI.

Provision of safety measures. Third, given an objective of providing, updating, and enforcing safety measures, another ideal characteristic is that the organization provides safety measures.[200] In evaluating whether a comparable treaty provides "safety measures," this Article assesses whether the treaty incorporates substantive requirements or procedural processes that mitigate the specific risks the treaty's objective seeks to address.

Factors from AI experts (expert factors). Further, another useful characteristic of an AI governance regime is possessing certain features noted by Markus Anderljung and other AI experts, such as (a) including expert-driven, multi-stakeholder processes; (b) being capable of rapid iteration; and (c) relying on technically informed processes.[201] 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. Another pertinent characteristic for an international legal regime is that it produces 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 (1) trade and finance treaties and (2) treaties incorporating enforcement mechanisms.[202] To supplement these quantitative analyses, Hoffman et al. provide qualitative data of other studies evaluating treaties for whether they produced their intended effects.[203] To determine if a given treaty possesses the "intended effects" characteristic, this Article uses both the meta-

---

[199] *See supra* Part I.B (describing VCPS and the Montreal Protocol). It also discusses the Cartagena Protocol, which bases its regulatory scope on a term's definition but also incorporates risk assessments.

[200] Harris et al., *Governance of Dual-Use Technologies*, *supra* note 78, at 6 (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* Anderljung et al., *Frontier AI Regulation*, *supra* note 40, at 23–29 (discussing potential safety standards for frontier AI).

[201] Anderljung et al., *Frontier AI Regulation*, *supra* note 40, at 3, 16, 21.

[202] Steven J. Hoffman et al., *International Treaties Have Mostly Failed to Produce Their Intended Effects*, 119 PROC. NAT'L ACAD. SCI. 1, 3 (2022), https://www.pnas.org/doi/full/10.1073/pnas.2122854119#t02.

[203] *Id.* at tbl. S9.

analysis determination for a specific treaty and the qualitative studies cited in the Hoffman et al. appendix.

Despite the usefulness of the Hoffman et al. study, its limitations are also important to consider, suggesting the intended effects characteristic should not be dispositive or even have overly weighted importance. For instance, the meta-analysis is limited by the quantity and quality of its included primary studies, and it is possible that Hoffman et al.'s study missed relevant primary studies. 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.[204] Finally, Hoffman et al. encourage caution in interpreting causal interpretations of intended and unintended treaty impacts.[205]

Enforcement mechanisms. Finally, given the Hoffman et al. 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."[206] For example, specific sanctions or consequences include "financial sanctions on countries or expelling countries from treaty bodies and trade blocs."[207] The Article includes this factor as a stand-alone characteristic to account for treaties that may not have been included in the Hoffman et al. 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, Table 2 provides 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

[204] *Id.* at 8.
[205] *Id.* at 6.
[206] *Id.* at 7.
[207] *Id.* at 5.

Applicable Treaty or Institution	Dual-use nature	Risk-based governance scope	Provision of safety measures	Expert Factors: (a) multi-stakeholder processes; (b) rapid iteration; (c) technically informed processes	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 International Court of Justice (ICJ) referral.
VCPS, 1971	Yes	Yes, art. 1(e).	Yes	(a): Yes (b): Yes (c): Yes	H: N/A QL: N/A	Yes, art. 31 provides for ICJ referral.
Convention Against Illicit Trafficking of Narcotic Drugs and Psychotropic Substances (CAIT), 1988	Yes	Yes. art. 1(n).	Yes	(a) Yes (b) Yes (c) No	H: N/A QL: N/A	Yes, art. 32.
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, [208] up to suspension by Parties.
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 (CBD).
Biological Weapons Convention (BTWC)	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)
Nuclear Non-Proliferation Treaty (NPT)	Yes	No	Yes, art. III(2).	(a): No (b): No (c): No	H: Yes QL: Not effective [209]	No

[208] U.N. ENV'T PROGRAMME, HANDBOOK FOR THE MONTREAL PROTOCOL ON SUBSTANCES THAT DEPLETE THE OZONE LAYER 800 (14th ed. 2020), https://ozone.unep.org/sites/default/files/Handbooks/MP-Handbook-2020-English.pdf.

[209] Hoffman et al. characterize a dissertation as stating the Nuclear Non-Proliferation Treaty (NPT) had the opposite effect of its intentions, but this statement is not necessarily true; the dissertation merely states the NPT may not be effective. *See* Taeshin Kwak, *Nuclear Non-*

Applicable Treaty or Institution	Dual-use nature	Risk-based governance scope	Provision of safety measures	Expert Factors: (a) multi-stakeholder processes; (b) rapid iteration; (c) technically informed processes	Intended Effects (Hoffman Meta-Analysis (H) & Qualitative List (QL))	Enforcement mechanism
Convention on Nuclear Safety [210]	Yes	No	Yes	(a): Yes (b): No (c): Yes	H: N/A QL: N/A	No
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	No	Yes, risk-based tiers based on categories, characteristics, and domestic legislation. art. 1	Yes. art. 4	(a): Yes (art. 10) (b): No, but art. 17 provides for amendments 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	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, this Part focuses 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: the Global Organization for High-Risk

Proliferation Regime Effectiveness: An Integrated Methodology for Analyzing Highly Enriched Uranium Production Scenarios at Gas Centrifuge Enrichment Plants 366–68 (2010) (Ph.D. dissertation, MIT), https://dspace.mit.edu/handle/1721.1/58084.

[210] 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. *See* James M. Acton, *On the Regulation of Dual-Use Nuclear Technology*, in GOVERNANCE OF DUAL-USE TECHNOLOGIES: THEORY AND PRACTICE (2016), https://www.amacad.org/publication/governance-dual-use-technologies-theory-and-practice/section/4 [https://perma.cc/78ZW-NPHX].

Artificial Intelligence (GOHAI).[211] In drawing on existing regulatory regimes to suggest components of GOHAI and a global regulatory system, this Article also aims to mitigate problems arising from legal transplants, meaning issues that arise when policymakers import elements of one legal regime to another despite mismatches in facts relating to the subject of regulation.

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,[212] GOHAI should include enforcement mechanisms that will ensure compliance or at least substantially increase the cost of noncompliance for Parties. "Enforcement mechanisms" 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."[213] For example, a treaty's enforcement mechanism may enable referral of disputes to an established judicial body, such as the

[211] Global Organization for High-Risk Artificial Intelligence 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 International Atomic Energy Agency (IAEA) has its Department of Management, Department of Technical Cooperation, Department of Nuclear Energy, Department of Nuclear Safety and Security, Department of Nuclear Sciences and Applications, and Department of Safeguards. *See* IAEA, *Organizational Structure*, https://www.iaea.org/about/organizational-structure [https://perma.cc/E8U7-7QGH].

[212] Hoffman et al., *supra* note 202, at 1.

[213] *Id.* at 7.

ICJ[214] or the Permanent Court of Arbitration.[215] Similarly, other treaties permit the establishment of an ad hoc arbitration panel or conciliation committee for the resolution of disputes.[216]

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.[217] 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 (UNSC).[218]

With this wide range of possibilities for GOHAI's enforcement mechanisms, treaty drafters should aim to ensure that 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 decisionmakers possess a competent level of scientific literacy, such as by including treaty provisions that require Parties to make efforts to train such key decisionmakers. Also, treaty drafters should ensure that binding dispute resolution mechanisms may be invoked by at least one party to a dispute and may not be evaded.[219] 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 noncombative resolution.[220]

---

[214] *See, e.g.*, Single Convention on Narcotic Drugs [hereinafter SCND] art. 48(2); VCPS, *supra* note 82, art. 31(2); Convention Against Illicit Trafficking in Narcotic Drugs and Psychotropic Substances [hereinafter 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, *supra* note 95, art. 20(2); IAEA Statute art. XVII.

[215] *See, e.g.*, Convention on International Trade in Endangered Species of Wild Fauna and Flora [CITES] art. XVIII(2).

[216] *See, e.g.*, SCND, *supra* note 214, art. 48(1); VCPS, *supra* note 82, art. 31(1); CAIT, *supra* note 214, art. 32(1); VCPOL, *supra* note 214, art. 11(3)(a); CBD, *supra* note 214, art. 27(3)(a); Basel Convention, *supra* note 95, art. 20(2); Convention on the Conservation of European Wildlife and Natural Habitats art. 18(2).

[217] *See, e.g.*, IAEA Statute, *supra* note 214, art. XIX.

[218] *See, e.g.*, Convention on the Prohibition of the Development, Production and Stockpiling of Bacteriological (Biological) and Toxin Weapons and on Their Destruction (BTWC) art. VI, Apr. 10, 1974, 1015 U.N.T.S. 163.

[219] *See, e.g.*, CITES, *supra* note 215, art. XVIII(2) (requiring mutual consent of Parties for arbitration).

[220] *See, e.g.*, VCPOL, *supra* note 214, art. 11 (allowing negotiation, arbitration, referral to the ICJ, or a conciliation commission).

2. Secretariat and Depositary

As the permanent administrative department of an international body, GOHAI will need a Secretariat. Secretariat functions may include organizing meetings of Parties, preparing reports—including annual status reports—and conducting scientific and technical studies.[221] GOHAI's Secretariat should be further empowered to provide Parties with information on private entities that could provide technical assistance on high-risk AI, especially considering the concentration of AI talent in the private sector.[222] Also, a catch-all provision permitting the Secretariat to complete any function assigned to it by the Parties would be useful.[223] Finally, treaty drafters should consider whether to establish a stand-alone Secretariat or to request the services of the UN Secretary-General as the treaty Secretariat.[224]

Further, the same entity that serves as the Secretariat may serve as the Depositary, which will be entrusted with the treaty itself.[225] A Depositary's functions may include informing Parties of the date a treaty will come into force, as well as providing notifications of any withdrawals, amendments, and updates to any annexes.[226]

3. Financial Mechanisms

For the financing of GOHAI and its initiatives, treaty drafters should seek to implement predictive, robust funding options. Looking to existing practices, treaties integrated into UN bodies may delegate determination of expenses to the UN General Assembly.[227] Similarly, Parties may have a subgroup of Parties propose an initial annual budget that must be subsequently approved by a two-thirds majority of all Parties

[221] For other descriptions of the functions of Secretariats, see SCND, *supra* note 214, arts. 16, 18; VCPOL, *supra* note 214, art. 7; Montreal Protocol, *supra* note 78, art. 7; CBD, *supra* note 214, art. 24; Cartagena Protocol, *supra* note 89, art. 31; CITES, *supra* note 215, art. XII; Basel Convention, *supra* note 95, art. 16.
[222] *See, e.g.*, Basel Convention, *supra* note 95, art. 16(h).
[223] *See, e.g.*, CITES, *supra* note 215, art. XII(2)(i).
[224] *See, e.g.*, SCND, *supra* note 214, art. 16.
[225] *See, e.g.*, CBD, *supra* note 214, art. 41; Basel Convention, *supra* note 95, art. 28; VCPOL, *supra* note 214, art. 20.
[226] *See, e.g.*, VCPOL, *supra* note 214, art. 20.
[227] *See, e.g.*, VCPS, *supra* note 82, art. 24.

to a treaty.[228] To supplement these budgetary mechanisms, treaty bodies may also establish trust funds to finance treaty initiatives.[229]

Moreover, financial mechanisms may support 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.[230] 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 Party to GOHAI should be required to establish designated Management and Scientific Authorities.[231]

For instance, CITES requires all Parties to designate a Management and Scientific Authority for specified objectives.[232] Through regulatory tools, including import and export controls, CITES seeks to protect endangered animal and plant species from overexploitation from international trade.[233] According to Article IX of CITES, each Party must designate a Management Authority "competent to grant permits or certificates on behalf of that Party."[234] Article IX also requires each Party to establish a Scientific Authority, which, among other responsibilities, must

---

[228] *See, e.g.*, IAEA Statute, *supra* note 214, art. XIV.
[229] *See, e.g.*, UNITED NATIONS ENVIRONMENT PROGRAMME, *supra* note 208, at 652.
[230] *See, e.g.*, Montreal Protocol, *supra* note 78, art. 10.
[231] *See, e.g.*, CITES, *supra* note 215, art. IX. *See also* VCPS, *supra* note 82, arts. 5(b)–(c), 6 (establishing functions for the consular, including the development of scientific relationships); SCND, *supra* note 214, art. 17 (permitting a special administration for the purposes of applying the treaty's provisions); Basel Convention, *supra* note 95, art. 5 (permitting the designation and establishment of competent authorities).
[232] Taking the United States as an example, the Division of Management Authority and the Division of Scientific Authority of the Fish and Wildlife Service in the Department of the Interior (DOI) exercise these responsibilities. Pervaze A. Sheikh & M. Lynne Corn, Cong. Rsch. Serv., RL32751, *The Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES)* (2016) [hereinafter CITES CRS Report], https://www.everycrsreport.com/files/20160921_RL32751_adf3a8a6b5526194b1439c7baddcdef19bda1e8c.html#Content [https://perma.cc/N2QY-3HU6].
[233] CITES, *supra* note 215, at Preamble.
[234] *Id.* art. IX(1)(a). For a concise summary of a Management Authority's responsibilities, see CITES CRS Report, *supra* note 232.

determine whether the granting of an import or export permit will have a harmful effect on the conservation status of a species.[235]

Similar to the CITES system, GOHAI could mandate a global network of Management and Scientific Authorities. Under one approach, GOHAI could oversee a global licensing system where each Party outlaws certain high-risk AI by default and only permits their release or training with a license.[236] Under this framework, each Party's 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 to grant a license to a model or to impose further conditions on its training or release. It could also determine 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, the framework should establish a global, impartial, and apolitical body of experts as a separate body from GOHAI's policymaking and advocacy entities. For example, SCND, the VCPS, and the CAIT (together, the Drug Treaties) partially rely on the WHO as a stand-alone body for some core functions.[237] As a UN agency, WHO aims "to promote health, keep the world safe and serve the vulnerable—so everyone, everywhere can attain the highest level of health."[238] To ensure the safety of the world

[235] CITES, *supra* note 215, art. IX(1)(b); *see generally* CITES CRS Report, *supra* note 232 (discussing when permits can be issued, along with some specific requirements for acquiring permits).

[236] 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, 1136 (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, 1373 (2021) (discussing the recent expansion of the First Amendment's freedom of speech protection to code).

[237] *See, e.g.*, SCND, *supra* note 214, art. 3(1); VCPS, *supra* note 82, art. 2(1); CAIT, *supra* note 214, art. 14(4).

[238] *About WHO*, WORLD HEALTH ORG., https://www.who.int/about [https://perma.cc/58CF-49WH].

from AI threats, the UN could establish a new agency: the World Artificial Intelligence Organization (WAIO).

Alternatively, the treaty creating GOHAI could create a subsidiary body serving a similar function and composed of a global, impartial group of experts. For example, the CBD and the Cartagena Protocol rely on a Subsidiary Body on Scientific, Technical, and Technological Advice (SBSTTA) to provide Parties with advice related to the implementation of the CBD and the Cartagena Protocol.[239] However, the SBSTTA meets more intermittently than the more permanent WHO.[240] Given the significant threats posed by AI, treaty drafters should favor the creation of a permanent organization, like WHO, as permanent oversight is superior to temporary oversight for a long-term threat of this nature.

Finally, to further cooperation for AI safety research, WAIO could serve as a clearing-house to facilitate the exchange of information, assist with implementation of the treaty, and make non-dangerous information freely available.[241] However, further research should be conducted to determine the most efficient approach to AI safety research, considering existing initiatives, such as the Global Partnership on Artificial Intelligence.[242] Treaties may further establish scientific centers for research and education relating to their treaty objective.[243] Drafters should also consider methods of integrating AI experts working in the private sector into these initiatives, given 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*

---

[239] CBD, *supra* note 214, art. 25; Cartagena Protocol, *supra* note 89, art. 30.
[240] *See Subsidiary Body on Scientific Technical and Technological Advice (SBSTTA),* CONVENTION ON BIOLOGICAL DIVERSITY, https://www.cbd.int/sbstta/#:~:text=Article%2025%20of%20the%20Convention,other%20subsidiary%20bodies%2C%20with%20timely [https://perma.cc/2XAH-SYHU] (noting that the SBSTTA has only met 26 times).
[241] Cartagena Protocol, *supra* note 89, art. 20; *see also* BIOSAFETY CLEARING-HOUSE, *Home*, https://bch.cbd.int/en/ [https://perma.cc/P77E-PAGW] (describing itself as "an online platform for exchanging information on [LMOs] and a key tool for facilitating the implementation of the Cartagena Protocol on Biosafety").
[242] OECD, *supra* note 18.
[243] *See, e.g.*, SCND, *supra* note 214, art. 38 bis.

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 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 impartiality.

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.[244] 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.[245] The Board's mandate covers tasks that hinge on its impartiality and credibility, including administering an estimate system of drug requirements,[246] administering the statistical returns system of various uses of regulated drugs and related items,[247] and preparing reports on its work.[248]

The Drug Treaties also task the Board with key advocacy functions aimed at ensuring the treaties' effectiveness.[249] For instance, in situations where a state becomes 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 consultation process with the state; request remedial measures; propose and offer assistance in a study of the matter; and, if those measures fail, call the attention of all states and the global public to the matter.[250]

---

[244] *See, e.g.*, *id.* art. 14; VCPS, *supra* note 82, art. 19; CAIT, *supra* note 214, art. 22.
[245] SCND, *supra* note 214, arts. 9–10.
[246] *Id.* arts. 12, 19.
[247] *Id.* arts. 13, 20.
[248] *Id.* art. 15; VCPS, *supra* note 82, art. 18. The SCND also states that Board members should possess "impartiality and disinterestedness." SCND, *supra* note 214, art. 9(2).
[249] SCND, *supra* note 214, art. 14; VCPS, *supra* note 82, art. 19; CAIT, *supra* note 214, art. 22. *See also* SCND, *supra* note 214, art. 14 bis (permitting the Board to recommend technical or financial assistance to a state, as long as said state agrees).
[250] SCND, *supra* note 214, art. 14. This process also requires the concern to arise "without any [State] failure in implementing the provisions of the [SCND]." SCND, *supra* note 214, art. 14(1)(a). *See also* VCPS, *supra* note 82, art. 19 (providing for a similar process); CAIT, *supra* note 214, art. 22 (same).

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 treaty's objectives. Under a process similar to that tasked to the International Narcotics Control Board by the Drug Treaties, the AI Control Board could work with individual Parties and encourage Parties and the international community to promote safe training before the release of prominent high-risk models. Similarly, if a Party becomes or risks becoming a center for the illicit training or production of high-risk models, the AI Control Board could work with that Party or draw the attention of the global community to the threat, 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 Parties and are empowered to oversee implementation of the treaty, including updating AI's risk-tier classifications.

For example, the Drug Treaties empower the Commission on Narcotic Drugs (the Drug Commission) and the ECOSOC to update the risk-tier classifications of substances (e.g., drugs) and their inputs (e.g., chemicals used to make such drugs).[251] ECOSOC is one of the six main organs of the UN and aims to advance three areas of sustainable development: economic, social, and environmental initiatives.[252] ECOSOC is composed of 54 states "elected for three-year terms by the [UN] General Assembly."[253] Subject to oversight by ECOSOC, the Drug Commission is the governing body of the United Nations Office on Drugs and Crime and oversees many regulatory activities detailed in the Drug Treaties.[254] The Drug Commission is composed of 53 states elected by

[251] SCND, *supra* note 214, arts. 7–8; VCPS, *supra* note 82, art. 17; CAIT, *supra* note 214, art. 21.
[252] *About Us*, U.N. ECON. & SOC. COUNCIL [hereinafter ECOSOC], https://ecosoc.un.org/en/about-us [https://perma.cc/3YFJ-529N].
[253] *Frequently Asked Questions*, ECOSOC, https://ecosoc.un.org/en/about-us/faq [https://perma.cc/R9ND-H8FU] (click arrow on "Where can I find the current membership?" box).
[254] *Commission on Narcotic Drugs*, UN OFF. ON DRUGS & CRIME, https://www.unodc.org/unodc/en/commissions/CND/index.html [https://perma.cc/NUY4-F44F].

ECOSOC and "is chaired by a Bureau, including one member per Regional Group." [255]

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 Drug Commission, while remaining subordinate and subject to the control of ECOSOC. Like the Drug Commission, CHAI could play a leading role in updating the risk-tier classifications of regulated substances and inputs—namely, AI models and ML chips. [256]

Alternatively, GOHAI could mimic the structure of the IAEA's policymaking bodies, particularly the General Conference and the Board of Governors. The General Conference consists of all IAEA states and has various powers, including electing members of the Board of Governors, suspending Parties, and approving amendments to the IAEA Statute. [257] The Board of Governors consists of 35 states that are geographically representative of the General Conference and also representative of the "most advanced [members] in the technology of atomic energy." [258] The Board of Governors has many powers, including: recommending to the General Conference a budget for the IAEA; considering applications for membership; approving safeguards agreements with individual states; approving the publication of the IAEA's safety standards; and appointing the Director General of the IAEA, with the approval of the General Conference. [259] To make decisions, both the General Conference and the Board of Governors generally require a two-thirds majority vote. [260]

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

---

[255] *Id.*

[256] *See* SCND, *supra* note 214, art. 8.

[257] IAEA Statute, *supra* note 214, arts. V(A), V(E)(1), V(E)(3), and V(E)(9). *See also* IAEA, *IAEA General Conference*, https://www.iaea.org/about/governance/general-conference [https://perma.cc/KA98-NNM5] (explaining the structure of the General Conference).

[258] IAEA Statute, *supra* note 214, art. VI(A); IAEA, *Board of Governors* [hereinafter IAEA, *Board of Governors*], https://www.iaea.org/about/governance/board-of-governors#:~:text=The%2035%20Board%20Members%20for,the%20Russian%20Federation%2C%20Saudi%20Arabia%2C [https://perma.cc/J4QQ-SFY5].

[259] IAEA, *Board of Governors*, *supra* note 258.

[260] IAEA Statute, *supra* note 214, arts. V(C), VI(E).

8. A Process for Updating High-Risk Models' Risk-Tier Classifications

Given high-risk AI's potential to have emergent capabilities or to pose previously undiscovered risks,[261] GOHAI should have a relatively rapid process for updating high-risk AI 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.[262] 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.[263]

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 or the Board may provide information to the UN Secretary-General of a possible needed change for a specific input's classification.[264] Second, the Secretary-General subsequently communicates this information to all states, the Drug Commission, and the Board, collecting further information as provided.[265] Third, based on this information, the Board decides whether the input is (1) "frequently used in the illicit manufacture of a narcotic drug or psychotropic substance," and whether (2) "the volume and extent of the illicit manufacture of a narcotic drug or psychotropic substance creates serious public health or social problems."[266] The Board's assessment is "determinative as to scientific matters."[267]

---

[261] *See* Hendrycks & Mazeika, *supra* note 9, at 5.
[262] CAIT, *supra* note 214, arts. 1(g), 1(t), 12(5).
[263] SCND, *supra* note 214, art. 1(u); VCPS, *supra* note 82, arts. 1(e), 1(g).
[264] CAIT, *supra* note 214, art. 12(2).
[265] *Id.* art. 12(3).
[266] *Id.* 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" and illicit purposes. *Id.*
[267] *Id.* art. 12(5).

Fourth, based on comments of states and the Board's determination, the Drug Commission decides in a two-thirds majority vote whether to enact the change to the input's classification.[268] Fifth, the Secretary-General communicates the Drug Commission's decision to all states and entities, and its decision is effective 180 days after this communication.[269] Sixth, if a state disagrees with this classification, it may appeal the Drug Commission's decision to ECOSOC, which may confirm or reverse the decision.[270]

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.[271] 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 or WHO, but not the Board, to notify the Secretary-General of a possible needed change.[272] In the second step, while all states and the Drug Commission receive communications, the WHO replaces the Board in receiving communications from the Secretary-General.[273] Further, pending the final decision of the Drug Commission as to the substance's classification, SCND permits the Drug Commission to provisionally consider that substance as a Schedule I substance—the highest risk tier—and states are required to apply control measures provisionally to this substance.[274] 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.[275] Fourth, based on the WHO's communications and any other relevant information, the Drug Commission decides whether to change a substance's classification.[276] Fifth, while not an option under SCND,

---

[268] *Id.*
[269] *Id.* art. 12(6).
[270] *Id.* art. 12(7).
[271] *See* SCND, *supra* note 214, art. 3; VCPS, *supra* note 82, art. 2.
[272] *Compare* SCND, *supra* note 214, art. 3(1), *and* VCPS, *supra* note 82, art. 2(1), *with* CAIT, *supra* note 214, art. 12(2).
[273] *Compare* SCND, *supra* note 214, art. 3(2), *and* VCPS, *supra* note 82, art. 2(2), *with* CAIT, *supra* note 214, art. 12(3).
[274] SCND, *supra* note 214, art. 3(3)(ii). The VCPS takes a less binding approach: states may "examine . . . the possibility" of such a provisional reclassification of a substance. VCPS, *supra* note 82, art. 2(3).
[275] SCND, *supra* note 214, arts. 3(3)(iii), 3(4), 3(5); VCPS, *supra* note 82, art. 2(4). For psychotropic drugs, the WHO's determinations are "determinative as to medical and scientific matters." VCPS, *supra* note 82, art. 2(5).
[276] SCND, *supra* note 214, art. 3(6); VCPS, *supra* note 82, arts. 2(5), 2(6).

VCPS permits states to opt out of certain aspects of the new classification, subject to ongoing minimal control measure obligations.[277] Sixth, SCND permits states 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.[278] Under both SCND and VCPS, ECOSOC may then "confirm, alter or reverse" the Drug Commission's decision, and such determination will be final.[279]

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 at least two tiers for ML chips' classification and at least two tiers of risk classification for high-risk AI models.

First, for both processes, any Party, its Management and Scientific Authorities, WAIO, or the AI Control Board could be empowered to notify or provide relevant information to the Secretariat regarding a high-risk model.[280] Among these entities, there are few compelling reasons to limit which entity may inform the Secretariat of such a concern: more scrutiny of dangerous models is generally desirable, and these entities are less likely than many others 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 entities,[281] considering the possibility that some information should not be widely communicated. For example, if the AI Control Board is made aware of a high-risk model capable of designing bioweapons, the Secretariat should not publicize that information but merely communicate with those entities that need to know that information to mitigate that risk, such as the AI model or system's developer, state of origin, and possibly states exposed to harm by the model or program.

---

[277] VCPS, *supra* note 82, art. 2(7).
[278] SCND, *supra* note 214, art. 3(8); VCPS, *supra* note 82, art. 2(8)(a).
[279] SCND, *supra* note 214, art. 3(8)(c); VCPS, *supra* note 82, art. 2(8)(c).
[280] *See* SCND, *supra* note 214, art. 3(1); VCPS, *supra* note 82, art. 2(1); CAIT, *supra* note 214, art. 12(2).
[281] *See* SCND, *supra* note 214, art. 3(2); VCPS, *supra* note 82, art. 2(2); CAIT, *supra* note 214, art. 12(3).

Third, for ML chips' classification, the AI Control Board could determine which ML chips warrant placement within respective risk tiers given an ML chip's capabilities.[282] 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 AI.[283] Similarly, for high-risk models' classification, WAIO could conduct a risk assessment as to the models' capabilities, and its finding should be determinative as to scientific and technological matters.[284] Pending the AI Control Board's or WAIO's decisions, CHAI should be empowered to provisionally classify an ML chip or AI model into a given Schedule, and this classification should be binding on Parties to apply the tier's respective safety measures.[285]

Fourth, for both processes, CHAI should be empowered to make an informed decision on the classification of an 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 AI, the classification should be applied provisionally by the Parties and take permanent effect as soon as reasonably possible, such as after 90 days.[286] Seventh, within a reasonable, short period, Parties should be permitted to appeal CHAI's decision to ECOSOC, which should be enabled to reverse, alter, or confirm CHAI's decision.[287]

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

9. A Representative Body of All Parties, Including Possible Subsidiary Bodies

[282] *Cf.* CAIT, *supra* note 214, art. 12(4) (authorizing the International Narcotics Control Board to take into account the capabilities of the substance in its scheduling decision).
[283] *See* Shavit, *supra* note 12, at 8–9.
[284] *Cf.* VCPS, *supra* note 82, art. 2(5) (declaring that the General Commission's assessment "shall be determinative as to medical and scientific matters").
[285] *Cf.* SCND, *supra* note 214, art. 3(3)(ii) (granting the General Commission the power to provisionally classify a substance in Schedule I, binding all Parties to matching regulations).
[286] *Cf.* SCND, *supra* note 214, art. 3(8) (limiting the period of review for the General Commission's scheduling decisions to ninety days).
[287] SCND, *supra* note 214, art. 3(8)(c) (confirming that ECOSOC has the authority to "confirm, alter or reverse" any decision of the General Commission); VCPS, *supra* note 82, art. 2(8)(c) (same).

Finally, meetings of the Parties should occur on a consistent, frequent basis and possess the ability to convene without delay in case of emergency developments. While this Article has 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 into the UN.[288] Possible functions of the COP include the ability to vote on amendments and to review the effectiveness of the treaty, such as Party compliance with oversight of safety practices by labs in their jurisdictions.[289]

III. Global Governance for Accountability

Part II defined 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. It argues for an objective of maintaining accountability for Parties and developers of high-risk AI to address misuse risks. After presenting existential risks from high-risk AI misuse, this Part explores precedents and suggests possible legal models and adaptations for a global system to ensure the accountability of states and actors working with high-risk AI.

By accountability, this Article refers to mechanisms that enable transparency, oversight, complaints, and enforcement of the treaty's objectives.[290] By transparency, this Article means information-sharing or reporting mechanisms imposed on Parties that would require them to share information with the international community, other states, or particular entities within an international organization.[291] By oversight, this Article refers both to mechanisms that enable active monitoring of regulated substances under a treaty and of Parties' compliance with treaty

[288] VCPOL, *supra* note 214, art. 6; CBD, *supra* note 214, art. 23; CITES, *supra* note 215, art. XI; Basel Convention, *supra* note 95, art. 15. The Conference of Parties model may also be compared to the IAEA's General Conference. *See* IAEA Statute, *supra* note 214, art. V.
[289] CITES CRS Report, *supra* note 232.
[290] Steven J. Hoffman et al., *supra* note 202, at 1.
[291] This definition is slightly more inclusive than that in the Hoffman et al. study, which defines transparency mechanisms as those that "enable information to be shared about countries with observers through regular reporting and information aggregation." *Id.* at 7.

obligations.[292] Complaint mechanisms permit "grievances attributable to a country to be processed and adjudicated through country or secretariat complaint mechanisms."[293] Finally, while this Article has previously used the Hoffman et al. definition of enforcement,[294] in this Part, this Article focuses solely on enforcement mechanisms that enable the possibility of sanctions or consequences delivered by a court or other authority on a Party for a failure to control private entities under its jurisdiction that have acted against a treaty's objectives.

While the Hoffman et al. study found that transparency, complaint, and oversight mechanisms were not necessarily associated with a treaty's effectiveness,[295] this Article suggests that such mechanisms may still be useful for global governance of high-risk AI. Given the nature of AI models as partly software that are 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 AI's 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 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, this Article's definitions of transparency, oversight, and enforcement vary from those used in the Hoffman et al. study, so the study results may not be fully applicable to the used definitions.

A. Misuse of High-Risk AI

This Part explains ways that high-risk AI may be misused by malicious or reckless actors. It argues that these misuse cases pose

---

[292] This definition is also more inclusive than that in the Hoffman et al. study, which defines oversight mechanisms as those that "build on transparency by actively monitoring and evaluating countries through standard setting and implementation review." *Id.*

[293] *Id.*

[294] *See supra* Part II.D (noting the possibility of a specific sanction or consequence by a court or other legal authority).

[295] Hoffman et al., *supra* note 202, at 6.

existential risks to humanity and necessitate a global governance regime to promote accountability and enable 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 AI that autonomously seek to kill billions of humans.

First, misuse of increasingly capable and accessible high-risk AI may increase the risk of bioterrorism. For instance, certain foundation models may enable nonexpert users to create deadly pathogens with step-by-step instructions.[296] These risks are augmented by the potential diffusion of similar foundation models, a scenario that would increase the number of people capable of creating such pathogens from a current number of about 30,000 experts to any nonexpert with access to such a model.[297] Along with an AI's ability to enhance the probability of the creation of a known pathogen, current AI are already capable of designing novel biological or chemical weapons.[298]

In a prominent example of this danger, a pharmaceutical company's research team discovered that its AI designed to penalize toxicity and bioactivity could easily be repurposed to reward these metrics, creating a tool for the design of new biological or chemical weapons.[299] 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.[300] This area is poorly regulated, and researchers largely have independent discretion over whether to proceed with such dangerous experiments.[301]

Second, along with AI-related bioterrorism risks, an additional misuse case would involve a human willingly or recklessly developing or releasing unaligned AI or AI tailored to kill. Given the likelihood of AGI being unaligned,[302] its intentional or accidental release could result in

---

[296] HENDRYCKS, MAZEIKA, & WOODSIDE, *supra* note 9, at 7; U.K. DEP'T FOR SCI., *supra* note 123, at 79.

[297] *See* Kevin M. Esvelt, *Delay, Detect, Defend: Preparing for a Future in Which Thousands Can Release New Pandemics*, GENEVA PAPER RESEARCH SERIES, No. 29/22, Nov. 2022 at 11 https://dam.gcsp.ch/files/doc/gcsp-geneva-paper-29-22.

[298] HENDRYCKS, MAZEIKA, & WOODSIDE, *supra* note 9, at 7.

[299] Urbina et al., *supra* note 69, at 189.

[300] *Id.* at 189–90.

[301] *Id.* at 190.

[302] *See supra* Part II.A.

catastrophic results for humanity. Further, some leading AI leaders are declared "accelerationists," believing AI are destined to replace humans as a dominant species and that humans should nevertheless strive to quickly create AGIs, despite the unsolved alignment problem.[303]

Additionally, foundation models can be repurposed for dangerous objectives.[304] For instance, an anonymous programmer took advantage of an open-source project to bypass ChatGPT's safety filters,[305] creating ChaosGPT, an autonomous AI with goals to "[d]estroy humanity"; "[e]stablish global dominance"; "[c]ause chaos and destruction"; "[c]ontrol humanity through manipulation"; and "[a]ttain [i]mmortality."[306] Fortunately, ChaosGPT was nowhere near competent enough to accomplish its goals, but as AI progresses, the likelihood increases that a future AI would succeed in achieving these harmful goals.[307]

Given the unique characteristics of high-risk AI, accountability of key actors is needed to mitigate harmful results. At its core, a foundation model's parameters and other defining features are contained in software. Once this software is released to the public, it is nearly impossible to recall. Further, as demonstrated by ChaosGPT, attributing liability to an anonymous actor intent on using released AI for harmful purposes is difficult. Thus, one of the most effective areas of regulation may be targeting the entities developing such models and programs or the activities used to create high-risk models or programs with required safety practices. 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

---

[303] HENDRYCKS, MAZEIKA, & WOODSIDE, *supra* note 9, at 8 (describing accelerationist statements from Google co-founder Larry Page and eminent AI scientists Jürgen Schmidhuber and Richard Sutton). This Article does not comment on whether future AGIs should have rights; rather, it aims to emphasize that accelerationist viewpoints must recognize that a transition of control from humans to unaligned AGIs would likely involve enormous amounts of suffering.

[304] 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 1, 16 (2023), https://arxiv.org/pdf/2307.15043.pdf.

[305] Jose Antonio Lanz, *Meet Chaos-GPT: An AI Tool that Seeks to Destroy Humanity*, DECRYPT (Apr. 13, 2023), https://decrypt.co/126122/meet-chaos-gpt-ai-tool-destroy-humanity [https://perma.cc/S4NB-USSB]. In other words, the model behind ChatGPT was unaligned, and the safety filters meant to mitigate the model's flawed outputs failed. *Id.*

[306] *Id.*; HENDRYCKS, MAZEIKA, & WOODSIDE, *supra* note 9, at 8.

[307] HENDRYCKS, MAZEIKA, & WOODSIDE, *supra* note 9, at 8.

future, more capable ChaosGPT. This regulation is especially important given that the resources to create an initial, raw foundation model are scarcer than those needed for fine-tuning or model prompting.[308]

While domestic regulation is crucial for ensuring accountability for developers of foundation models, international regulation is a necessary complement, considering the impact that AI misuse would have on all states, regardless of its origins.[309] An international regime resulting in accountability of key actors—both private entities and states—building high-risk AI will enable emergency responses or regulation when necessary.

B. Accountability Mechanisms for High-Risk AI

This Part explores accountability mechanisms that would enable transparency, oversight, complaints, and enforcement of high-risk AI. It begins with a preliminary overview in Table 3 of such accountability mechanisms in existing international regulatory regimes. It then provides an in-depth overview of possible model regimes for accountability mechanisms, applying relevant components to a global governance regime for mitigating misuse of high-risk AI.

Table 3: Accountability Mechanisms in Existing Global Legal Regimes

Legal Regime	**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),	Yes, arts. 9(1), 9(3), 12(9)(a),	No	Yes, arts. 3, 4, 5, 6, 7, 15

[308] 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 supra* Part I.A.

[309] Shavit, *supra* note 12, at 4.

Legal Regime	Transparency mechanisms (*E.g.*, reporting or disclosure obligations)	Oversight mechanisms	Complaint mechanisms	Enforcement mechanisms for private entities
	12(10)(a), 12(11), 12(12), 20	16, 22		
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.)) [310]	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.
Montreal Protocol	Yes, arts. 2(5), 2(7), 7	Yes, arts. 4(1), 4(3), 4(6)	No	No
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 states.
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
BTWC	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)

[310] While not an agreement itself, INFCIRC/153 (Corr.) provides guidance for the IAEA Secretariat in the negotiation of comprehensive safeguards agreements with non-nuclear weapons states (NNWS), detailing desired structures and contents for such agreements. LAURA ROCKWOOD, IAEA, LEGAL FRAMEWORK FOR IAEA SAFEGUARDS 21 (2013), https://www.iaea.org/sites/default/files/16/12/legalframeworkforsafeguards.pdf.

1. Transparency Mechanisms

This Part 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 the regulation of 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 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.[311] Further, the IAEA Statute and the Cartagena Protocol mandate reports that share information relating to the effectiveness of projects or initiatives created by their respective treaty entities. These reports aim to advance knowledge relating to the peaceful use of nuclear energy and to LMOs, respectively.[312]

For global governance of high-risk AI, a foundational treaty could establish similar reporting requirements. For instance, 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 initiatives that result in the production of new knowledge—relating to the alignment problem, for example—Parties could be encouraged or mandated to provide such helpful information in reports or correspondences with WAIO, CHAI, the AI Control Board, or an entity similar to the Biosafety Clearing-House established in the Cartagena Protocol.[313]

Second, to regulate possible harmful substances, many treaties mandate that states report breaches or possible breaches of a treaty's objectives or of states' obligations. For instance, the Drug Treaties require reports of illicit trafficking of regulated drugs;[314] imports, exports, or

---

[311] SCND, *supra* note 214, arts. 18(1)(a)–(b) (mandating reports and texts of all laws and regulations be given to the UN Secretary-General); VCPS, *supra* note 82, art. 16(1) (same and covering "[s]ignificant developments" related to the treaty); CAIT, *supra* note 214, art. 20 (same); CITES, *supra* note 215, art. VIII(7)–(8) (same); CBD, *supra* note 214, art. 26 (same); Cartagena Protocol, *supra* note 89, art. 33 (same).

[312] IAEA Statute, *supra* note 214, art. VIII; Cartagena Protocol, *supra* note 89, art. 20(3).

[313] *See* Cartagena Protocol, *supra* note 89, art. 20.

[314] SCND, *supra* note 214, art. 18(1)(c); VCPS, *supra* note 82, art. 16(3).

transits of precursory substances that may result in the illicit manufacture of regulated substances;[315] and seizures of such precursory substances.[316] Similarly, the CBD and Cartagena Protocol encourage states 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 risks posed by any transboundary movement.[317]

Depending on which AI are regulated, a treaty for high-risk AI could mandate reporting possible harms or manifested harms that the treaty seeks to avoid. For instance, narrow models trained for specific contexts, such as for pharmaceutical research, pose the risk of enabling nonexperts to design chemical weapons.[318] If a company learns that one of its models or programs has been misused for such a purpose either by a researcher or a licensee of such a model, the treaty could require Parties to mandate that companies under their jurisdiction report such incidents to national regulatory authorities. These regulatory authorities could include the recommended Management and Scientific Authorities, the WAIO, and possibly the WHO. These reports would need to consider mitigation methods for risks from misuse, as well as confidentiality protocols to share risk information with authorities to 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, such as ML chips, in each risk tier, as well as related components to such substances. CITES mandates such a system, requiring states to maintain records of the trade of endangered specimens included in Appendices I, II, and III.[319] The Drug Treaties require states to provide the Board with various estimates[320] and statistics[321] related to regulated drugs and related

[315] CAIT, *supra* note 214, art. 12(9)(c).
[316] *Id.* art. 12(a).
[317] CBD, *supra* note 214, art. 14; Cartagena Protocol, *supra* note 89, arts. 17, 25.
[318] *See supra* Part III.A; Urbina et al., *supra* note 69, at 189–90.
[319] CITES, *supra* note 215, art. VIII(6).
[320] SCND, *supra* note 214, art. 19(1) (requiring estimates for drug quantities used for medical and scientific purposes and for the manufacture of other drugs, as well as requiring an estimate for the area of land to be used for opium poppy cultivation, among others).
[321] *Id.* art. 20 (requiring statistics for quantities relating to the production or manufacture of drugs, the consumption of drugs, the imports and exports of drugs and poppy straw, and the seizures of drugs, among others); *see also id.* art. 27(2) (requiring Parties to separately furnish estimates and statistics for the amount of coca leaves used in preparing flavoring agents) and VCPS, *supra* note 82, art. 16(4)–(5) (requiring annual statistical reports for each schedule, as well as requiring the furnishment of supplementary statistical information upon request).

substances. Finally, with its quantitative approach to regulation, the Montreal Protocol requires initial reporting and annual reporting of regulated substances—both to establish a baseline regulatory amount and to compare progress against that baseline.[322]

To effectively regulate certain high-risk AI, Yonadav Shavit and Mauricio Baker have suggested adapting systems to include the reporting of certain materials or locations that could 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,[323] Parties could be required to report a baseline amount of noncompliant chips, as well as to annually report quantities of compliant chips, the number of high-quality ML chip fabrication facilities (fabs), data centers, storage facilities, and elimination facilities.[324] 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 other dangerous contexts poses high risks.

Fourth, in the context of import and export controls of regulated substances, many treaties require states to report certain movements of regulated substances when crossing state borders. The Drug Treaties, for example, require or permit states to notify other states of exports of regulated drugs or precursory substances, as well as of prohibited imports of regulated drugs.[325]

For high-risk AI, similar to the Biden administration's trade controls designed to limit China's access to AI-related materials, a global treaty could impose similar reporting requirements when regulated materials are moved internationally, setting a foundation for a global oversight system.[326] 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

[322] *See* Montreal Protocol, *supra* note 78, art. 7.
[323] Shavit, *supra* note 12, at 2.
[324] Mauricio Baker, *Nuclear Arms Control Verification and Lessons for AI Treaties*, ARXIV 1, 37–42 (2023), https://arxiv.org/pdf/2304.04123.pdf.
[325] VCPS, *supra* note 82, arts. 12(2), 13; CAIT, *supra* note 214, art. 12(10).
[326] For such a suggested global AI oversight system, see Baker, *supra* note 324, at 37–42.

equipment.[327] While the second Trump administration rescinded certain Biden administration AI rules, many of the Biden-era export controls remain in place.[328] 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, given the likely widespread diffusion of high-risk AI, Parties could be requested to report needed regulations or changes to the risk tier for a given AI model.[329]

2. Oversight Mechanisms

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

Generally, many treaties pursue oversight of regulated substances by having states agree to implement domestic measures that will be uniformly applied on a global scale, resulting in a reduced risk of harm. For example, the VCPS requires states 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."[330] The SCND encourages states to require licensing of individuals working with regulated drugs, mandating that such persons "have adequate qualifications" to comply with laws and regulations enacted in compliance with the convention.[331]

---

[327] GREGORY C. ALLEN, CTR. FOR STRATEGIC & INT'L STUD., CHOKING OFF CHINA'S ACCESS TO THE FUTURE OF AI 2 (2022), https://csis-website-prod.s3.amazonaws.com/s3fs-public/2023-04/221011_Allen_China_AccesstoAI.pdf?VersionId=V2RCmPjpR8nQOybvB2LmNIu1Yx6RlYvA.

[328] STEPTOE LLP, *Trump Administration Charts New Path on AI Export Controls with Significant New Guidance and Rescission of Diffusion Rule*, INT'L COMPLIANCE BLOG (May 19, 2025), https://www.steptoe.com/en/news-publications/international-compliance-blog/trump-administration-charts-new-path-on-ai-export-controls-with-significant-new-guidance-and-rescission-of-diffusion-rule.html [https://perma.cc/6GW5-LJD9].

[329] *Cf.* VCPS, *supra* note 82, arts. 2–3 (requiring Parties to institute certain restrictions on substances depending on the risk tier they fall under and requiring Parties to categorize substances to risk tier appropriately); *see also* the process suggested in *supra* Part II.E.8.

[330] VCPS, *supra* note 82, art. 15.

[331] SCND, *supra* note 214, art. 34(a).

Further, both Parties and private entities could be required to abide by recordkeeping requirements.[332]

Global agreements for high-risk AI could encourage similar, uniform domestic measures mandated in previous treaties. Like the SCND's encouragement for 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 Party recordkeeping system as explained in Part III.B.1, a state may need to require private entities to establish their own recordkeeping 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 of hardware, such as ML chips.[333] As explained in Part II.E.4, CITES imposes a global system of import and export controls partially through reliance on states' individual Management and Scientific Authorities to grant import or export permits and certificates.[334]

Based loosely on CITES, a global trade oversight system for materials used to train and build high-risk AI 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 Biden-era trade restrictions toward China.[335] In such a system, 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 to permit the import or export of such ML chips.[336]

Finally, given the breadth of global oversight of nuclear-related materials and of Parties dealing with such materials,[337] many have

---

[332] For such requirements for Parties, see *supra* Part III.B.1.

[333] *See supra* Part III.B.1 for reporting mechanisms in the trade context.

[334] *See supra* Part II.E.4; CITES, *supra* note 215, art. IX.

[335] *See* ALLEN, *supra* note 327, at 2.

[336] *Cf. supra* Part II.E.4, where the author recommends that Management and Scientific Authorities could license AI models' use within a state, not in the international 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.

[337] Baker, *supra* note 324, at 25.

suggested that the global nuclear regulatory system may offer guidance for the regulation of high-risk AI, which likely requires similar robust oversight.[338] 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, such as promoting and restricting the proliferation of nuclear technologies. 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;[339] to increase the proliferation of nuclear technology to guard against lack of reliable energy;[340] 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.[341] In the nuclear regulatory context, "safeguards" are oversight activities through which the IAEA may verify that a state is not using nuclear materials to create unauthorized nuclear weapons.[342]

This Article has focused primarily on objectives relating to safety, not objectives similar to those in the nuclear context, such as proliferation of capable AI or the prevention of unaligned AI's proliferation. Even with this focus on safety, however, comparisons to the nuclear safety regime may yield limited results. The IAEA focuses primarily on oversight of tangible materials used to create nuclear weapons. In the AI context, while many tangible materials are important for the creation and functioning of AI models, such as ML chips and data centers, a finished model is largely intangible. Also, with nuclear materials, there are two possible outcomes for their use: a state uses them for peaceful purposes, such as energy, or a state uses them to create a nuclear weapon. For high-risk AI, use outcomes are not so clear; an entity may attempt to use materials to train a capable, safe AI that is later discovered to be dangerous.[343]

---

[338] For an alternative method of minimal oversight of states to a treaty, see CAIT, *supra* note 214, art. 22.

[339] Acton, *supra* note 210 (declaring that the objective of security efforts is to "prevent the unauthorized possession of nuclear material").

[340] *See* IAEA Statute, *supra* note 214, art. II.

[341] *See Id.* art. III(A)(5).

[342] IAEA, *IAEA Safeguards Overview*, https://www.iaea.org/publications/factsheets/iaea-safeguards-overview (last visited Apr. 17, 2026).

[343] *See supra* Part II.A.

Thus, rather than focus on the possible implementation and workings of oversight mechanisms for high-risk AI based on the nuclear safety regime's materials-based approach, this Article briefly explores possible secondary objectives of similar oversight regimes for high-risk AI, drawing on established risks and recognizing that a materials-based oversight system may be limited in its practicality.[344]

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.[345] Focusing on the former risk, misuse could be defined relative to noncompliance with specific model restrictions in various manners, such as those based on total training compute, properties of training data, properties of hyperparameters, performance-based benchmarks, bans, or post-training safety mitigations.[346] 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 or other modifications helpful for monitoring compliance with specified requirements. Regardless of the objective, a monitoring system would require the establishment of a permanent staff for global inspections, possibly operating under GOHAI's direction.[347]

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.[348]

---

[344] For such possible oversight mechanisms for high-risk AI and overviews of the agreements composing the global nuclear safety regime, see Shavit, *supra* note 12, at 4–6; Baker, *supra* note 324, at 12–19, 22–25; ROCKWOOD, *supra* note 310, at 4–10, 27–30; IAEA, *Basics of IAEA Safeguards*, https://www.iaea.org/topics/basics-of-iaea-safeguards [https://perma.cc/H7S5-FGXC]; IAEA, *More on Safeguards Agreements*, https://www.iaea.org/topics/safeguards-legal-framework/more-on-safeguards-agreements [https://perma.cc/78BE-PFFC].

[345] For a focus on the latter, see *supra* Part II.

[346] Shavit, *supra* note 12, at 4–5.

[347] *See* IAEA Statute, *supra* note 214, art. VII(C) (establishing a similar staff for the IAEA).

[348] *Compare* Treaty on the NPT, art. III.1 (requiring NNWSs to accept safeguards), *with id.* art. III.2 (requiring State Parties to withhold providing equipment or material to any NNWS unless the material is subject to IAEA safeguards); *see also* ROCKWOOD, *supra* note 310, at 4–5 (elaborating on this comparison).

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."[349] Two separate groups may emerge in the context of regulation for AI-related materials—possibly, (1) the U.S. and its allies, and (2) non-allied states, such as China and Russia. However, further research is needed on methods for ensuring safety for all of humanity and whether that would encompass encouraging such a two-tiered system.

3. Complaints

For complaint mechanisms, despite the treaty's lack of institutional support entities, BTWC offers a potential model, permitting the lodging of complaints for grievances attributable to a state with the UNSC.[350] The BTWC has been critiqued for its failure to prevent the development of prominent biological weapons programs, such as that of the former Soviet Union.[351] Notably, the BTWC 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.[352] While valid, critiques of the BTWC 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

[349] Acton, *supra* note 210, arts. VI, IV.
[350] BTWC, *supra* note 218, art. VI.
[351] *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.
[352] UNITED NATIONS, OFF. FOR DISARMAMENT AFFS., *Implementation Support Unit*, https://disarmament.unoda.org/biological-weapons/implementation-support-unit/ [https://perma.cc/RUF5-SKNM]; *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/ (last visited Apr. 17, 2026).

powerful AI models with the capability to cause a deadly pandemic. Similarly, given the possible technical nature of some complaints, CHAI or the GOHAI Secretariat may need 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

Given that many private entities control or develop high-risk AI, an AI safety treaty should ensure there are sufficient enforcement mechanisms against private entities that undermine or act against the goals behind the treaty's objectives.[353] One possible way to apply such accountability to private entities is by requiring Parties to enact legislation or measures that hold private entities responsible through domestic civil or criminal law.[354] Many states operate according to a dualist perspective of international law, meaning that domestic legislation is needed to give effect to obligations undertaken in a treaty.[355] Thus, enacting legislation for a treaty could incorporate civil or criminal penalties for private entities. Parties that do not enact these civil or criminal penalties could be held responsible under enforcement mechanisms applicable to Parties.[356]

The Drug Treaties provide possible precedents of ensuring states 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 states to establish as criminal offenses situations where a person acts "intentionally" while committing any of a list of enumerated actions.[357] The SCND specifies actions which, among others, include "cultivation, production, manufacture,

---

[353] *See, e.g.*, Fabio Urbina et al., *supra* note 69, at 189. *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.

[354] Another approach outside the scope of this Article would be to explore the appropriateness and possible methods of providing for private entity-state arbitration to resolve disputes relating to high-risk AI.

[355] 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, at 1, https://supremecourt.uk/uploads/speech_170213_1d3b2802ba.pdf.

[356] *See supra* Part II.E.1.

[357] SCND, *supra* note 214, art. 36; VCPS, *supra* note 82, art. 22; CAIT, *supra* note 214, art. 3.

extraction, . . . importation and exportation of drugs contrary to the provisions of this Convention."[358]

Further, the Drug Treaties include provisions meant to ensure that states can effectively penalize violations of domestic laws and regulations that enact treaty provisions, as well as 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 states.[359] CAIT further requires states to adopt measures that enable the confiscation of proceeds from criminal offenses enumerated in article 3, paragraph 1, as well as for the confiscation of the regulated substances themselves.[360]

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 and the required mens rea. 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 Parties' domestic legislation, actions leading to the design of biological weapons with high-risk AI models could be widely prohibited absent a license requiring high safety standards and personnel preclearance. 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.[361]

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 states' domestic law. The Cartagena Protocol requires states to adopt civil measures and to consider criminal measures to prevent or penalize "transboundary movements of [LMOs] carried out in contravention of its domestic measures to implement this Protocol."[362]

---

[358] SCND, *supra* note 214, art. 36(1).
[359] CAIT, *supra* note 214, arts. 4, 6, 7.
[360] *Id.* art. 5.
[361] *Cf. id.* arts. 4–7 (establishing mechanisms for confiscation and extradition, where necessary, for adequate enforcement of the treaty's provisions as well as establishing jurisdiction to settle related disputes and legal assistance to Parties, if needed).
[362] Cartagena Protocol, *supra* note 89, art. 25.

The Nagoya-Kuala Lumpur Supplementary Protocol offers further tools for effectiveness, requiring any entity in control of an 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.[363]

Provisions like 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 AI should be obligated to report harms necessitating responsive mitigation measures.[364] 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 those situations.[365]

Finally, in the context of trade, CITES provides a model through which states agreed to criminalize certain trade of the subject of regulation: for CITES, these subjects are endangered species.[366] Notably, given the success of CITES toward its objectives,[367] its components may offer promising precedent for global AI governance.

For global governance of high-risk AI, an underinclusive approach could request Parties to impose criminal or civil liability for intentional trade of items in contravention of the treaty, such as certain ML chips, lithography equipment, and other components used in manufacturing ML chips.[368] As explored in Part III.B.2, this approach may be promising yet still be underinclusive, as many high-risk models may result from fine-tuning or model prompting of existing models.

Conclusion

[363] Nagoya-Kuala Lumpur Supplementary Protocol on Liability and Redress to the Cartagena Protocol art. 5.

[364] *Cf. id.* art. 5 (requiring both the entity and the competent state authority to report, evaluate, and redress harms).

[365] 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.

[366] CITES, *supra* note 215, art. VIII(1)–(2) (requiring states 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).

[367] *See* CITES CRS Report, *supra* note 232.
https://www.everycrsreport.com/files/20160921_RL32751_adf3a8a6b5526194b1439c7baddcdef19bda1e8c.html ("[N]o species listed under CITES within the last 30 years has gone extinct.").

[368] *See* Allen, *supra* note 327, at 1–2.

This Article argues 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. It argues for a framework of establishing global governance regimes based on (1) framing risks; (2) determining objectives intended to address such risks; and (3) 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 address 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 to facilitate decisions. With high-risk AI, regulators must eventually make such probability estimates before choosing among informed regulatory measures.

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

Fourth, based on analyses of existing 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, complaint, and enforcement mechanisms for Parties and private entities.

To meet the prominent challenges high-risk AI poses to humanity, further research is necessary. This Article does not explore the political challenges to constructing global governance regimes, nor does it focus on the best methods to ensure high-risk AI systems that are secure from persons or entities with ill intentions. 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.