Safety and Global Governance of Generative AI Report

WFEO-CEIT

Shenzhen Association for Science and Technology

Nov 2023

Table of Contents

Editors’ Note

This report is a compilation of diverse authorial perspectives, aimed at drawing public attention to the safety and governance issues surrounding the development of generative artificial intelligence and stimulating further thought. We recognize the rapid pace of advancement in this field, accompanied by numerous potential challenges and opportunities. The specific viewpoints expressed in this report represent only the individual authors and do not reflect the positions of the Standing Technical Committee on Engineering for Innovative Technologies (CEIT) of the World Federation of Engineering Organizations (WFEO) or the Shenzhen Association for Science and Technology. We emphasize the need for broad collaboration and ongoing dialogue both within and outside the industry to ensure that the advancements in generative AI benefit humanity and are managed appropriately within ethical and legal frameworks. Through this report, we hope to foster more exchanges and collaborations, jointly exploring the future of this cutting-edge technology.

Preface

AI Governance: Towards Better and Faster Development of AI to Accelerating the Implementation of the SDGs

GONG Ke (龚克)

Firstly, I want congratulate colleagues of the Standing Technical Committee on Engineering for Innovative Technologies (CEIT) of the World Federation of Engineering Organizations (WFEO) and the Shenzhen Association of Science and Technology, for their meaningful work of compiling this report - Report on Generative Artificial Intelligence Safety and Global Governance, in collaboration with the Shenzhen Association of Science and Technology. In this report, experts from academic and industry institutions of various countries provide insightful observations, reflections, and valuable experiences regarding AI governance from different perspectives. The crucial consensus points have been emphasized in the report, such as the need to establish a global, multi-stakeholder shared governance mechanism for AI, the importance of ethical considerations as the foundation of AI governance, and the prioritization of high-risk areas for establishing global governance standards. I also noticed some experts proposing concrete suggestions, for example, establishing high quality and open datasets or establishing international demonstration zones for AI value alignment and governance innovation. They call for global participation from AI companies and research institutions in the zone to collectively verify the scientific and operational aspects of alignment methods, standard specifications, governance tools, and data sharing mechanisms. This report serves as a valuable reference for policymakers, technology standard developers, and engineering managers involved in AI development and governance.

The release of this report coincides with the midpoint of the United Nations' 2030 Sustainable Development Agenda. At the recent UN Sustainable Development Goals (SDGs) Summit, world leaders unanimously called for intensified efforts to accelerate the implementation of SDGs. UN Secretary-General Mr. Antonio Guterres highlighted the significance of SDGs as embodying the hopes, dreams, rights, and expectations of people worldwide. However, only 15% of the goals are progressing as expected, and some are even regressing. Urgent action is needed to devise a global plan to salvage these goals. Mr. Antonio Guterres stressed actions in six key areas, one of which is "leveraging the opportunities of digital transformation." AI must become a significant driver in accelerating the achievement of SDGs.

Artificial intelligence is a revolutionary general-purpose technology that serves as the advanced productivity driving the Fourth Industrial Revolution and the digitalization of economies and societies. Whether at the global level (accelerating sustainable development transformation), regional and national levels (promoting regional and national energy transitions, promoting economic growth and employment), industry and enterprise levels (enhancing industry digital transformation, improving competitiveness and organizational efficiency), or individual levels (upskilling in one's career, and improving the convenience of daily life), AI holds immense potential. Therefore, AI governance should not hinder but rather facilitate its better and faster development to serve all people and leave no one behind.

The rapid development of AI, especially the recent advancements in generative AI, has brought us unprecedented experiences and excitement. However, it has also intensified concerns about AI safety and ethics, leading to societal anxiety. This highlights the importance and urgency of proper AI governance to ensure AI is under human control and good for all people. AI governance should be based on ethical principles, and the UNESCO Recommendations of AI is the first-ever UN document serving as the foundation for AI governance.

Given that emerging digital technologies like AI are inherently global, they ignore geopolitical as well as disciplinary boundaries. Their development and governance concern the common interests of all people regardless of nationality, ethnicity, religion, gender, age, and social status. The impacts they generate, whether positive or negative, transcend national borders, industrial sectors, and professional disciplines, having a global and overarching influence. Effective governance of AI must be a global, multi-stakeholder effort. While international organizations (such as the UN, G20, G7, OECD, EU, etc.) and governments, as well as AI companies, have taken actions in AI governance, there is still a need for broad and clear global consensus. As suggested by some report authors, wide-ranging dialogues involving diverse stakeholders should be convened within the UN framework to establish a mechanism for global pluralistic governance, akin to that for addressing climate change.

WFEO, driven by its mission to promote sustainable development through engineering, places significant emphasis on both promoting AI development and application to expedite the dual transformation (digital and sustainable transformation) and governing AI responsibly. In 2020, WFEO-CEIT released seven Principles for Responsible Conduct of Big Data and AI in Engineering on the 4th of March, the first World Engineering Day for Sustainable Development; in 2021, WFEO supported the UN Department of Economic and Social Affairs (UNDESA) and UN Office of the Secretary-General's Envoy on Technology in publishing the Resource Guide for AI Development Strategy. WFEO had also actively engaged in the consultation work for UNESCO's AI ethics recommendations. Recognizing that AI's development and application hinge on engineering and that only engineered AI can truly impact human production and life, the engineering community should play a vital and active role in AI governance.

From an engineering perspective, special emphasis should be laid on embedding AI ethical principles and legal provisions into technical standards. These standards should be testable and interoperable. To turn these principles and standards into reality, there should be a priority in developing technical means and tools that support governance requirements, such as privacy computing and ethical auditing technologies, etc.

WFEO also underscores the importance of comprehensive capacity building for AI development and governance, including engineering education. Taking practical steps to reduce the gap in digital capability related to AI should be a necessary part of global AI governance.

In conclusion, as leaders of the global engineering community, WFEO is willing to play an active role in global AI governance. We firmly believe that the unstoppable advancement and application of AI should be properly governed to ensure AI with human values for sustainable development. The governance of AI should aim to facilitate its better and faster development, maximizing its potential to serve the sustainable development of humanity and the Earth. We reiterate that effective governance of AI must be a global, multi-stakeholder, cooperative effort. Initiating broad dialogues within the UN framework to foster clear governance consensus and establishing a permanent mechanism (similar to climate agreements) should be the foundation for further advancing collective governance actions. We stress that AI governance should be ethical based governance and implementing UNESCO Recommendations of AI. We also note that the governance principles proposed and ongoing governance practices adopt a risk-based differentiated governance approach. Thus, we advocate for prioritizing on achieving a global consensus on AI risks. We further advocate for capacity building in AI development and governance, particularly in narrowing the gaps in AI capacity, as a crucial aspect of AI governance. And, in helping developing countries build AI capabilities, urgent and robust actions should be implemented.

Gong Ke (龚克), Immediate Past President, World Federation of Engineering Organizations (WFEO) 2019-2022, advisor to WFEO-CEIT.

Introduction

On October 18, 2023, Chinese President Xi Jinping announced China’s "Global AI Governance Initiative" during the keynote speech at the opening ceremony of the Third Belt and Road Forum for International Cooperation. The initiative was officially released by the Cyberspace Administration of China on the same day. It articulated China's positions and proposals regarding the development, security, and governance of artificial intelligence, expressing a willingness to communicate, exchange, and pragmatically cooperate with all parties on global AI governance, as well as promote the benefits of AI technology for humanity. On October 26, UN Secretary-General António Guterres announced the formal establishment of a new High-Level Advisory Body on Artificial Intelligence, comprising 39 experts from around the world and supporting international society strengthen governance by discussing the risks and opportunities presented by AI technology. Subsequently, on November 1, the inaugural Global AI Safety Summit was held at Bletchley Park in the United Kingdom. The EU and twenty-eight countries, including China and the United States, collectively signed the "Bletchley AI Safety Declaration," unanimously acknowledging the potential catastrophic risks posed by AI to humanity.

Against this backdrop, this report has gathered 29 essays from over 40 policymakers, entrepreneurs, experts, scholars, and engineers in the fields of AI governance, science and technology ethics, large model safety and alignment, AGI risks, etc. The essays engage in a dialogue on key topics related to the safety and global governance of generative artificial intelligence and are organized into five chapters based on themes.

Chapter 1: The Risks and Challenges of Generative AI

Experts looked at both short-term and long-term perspectives. Regarding immediate risks and challenges, first, there are privacy and security hazards associated with large language models (LLMs). Second, there are significant problems with the generation of false information and the problem of "hallucinations" by LLMs. Third, there are concerns about value biases in models and their lack of interpretability. Fourth, there are ethical and moral risks arising from the misuse of models. Lastly, there are intellectual property and legal regulatory issues stemming from AI applications.

Regarding long-term risks and challenges, first, AI may bring about major economic and social transformations that require coordinated responses. Secondly, it may overturn the existing international legal system and world order. Third, AI poses risks to the global commons, necessitating the establishment of international norms and regulations. Fourth, there are differences in values regarding AI among different countries and cultures. Fifth, strong AI may potentially escape human control, creating catastrophic risks.

In summary, short-term risks are more concentrated at the level of individual models and applications, while long-term risks and challenges are more related to the overall development direction of AI technology and its societal impact. Regardless of the timeframe, international cooperation and the formulation of ethical standards are crucial for addressing the risks posed by AI.

Chapter 2: Global Governance Approaches for Generative AI

The focus and priorities of experts are summarized below. The first focus is global cooperation and coordination, including emphasizing the necessity of international cooperation in AI governance, urging the prompt initiation of multilateral coordination and cooperation processes, and promoting broad participation of countries in governance. The second topic identified is risk identification and management, concentrating on potential shared, large-scale, and high-risk hazards posed by AI systems. Third, experts prioritize ethics and transparency, stressing the importance of ethical principles and transparency in the design, development, and deployment of AI. Additionally, they call for balancing technological development with safety and ensuring safety and compliance at the same time as promoting technological innovation.

In response to these priorities, experts proposed several policy recommendations. They call for working together with multinational organizations and international society, pushing for the creation of international AI organizations to ensure the implementation of international supervisory standards since global participation and cooperation is needed to identify and mitigate AI risks. They also propose developing risk warning and response mechanisms including post-hoc supervision and review and preventive strategies, in order to ensure system safety, security, and reliability. Their third proposal is the creation of third-party assessment mechanisms, since independent expert third-party assessments can complement internal evaluations, providing a robust safety net. They additionally support construction of interoperable compliance systems, advocating for the standardization and alignment of governance rules across different countries. Lastly, they call for instituting international agreements to share the results and benefits of AI globally.

Overall, there is an urgent need for a unified global perspective to reach consensus on principles and policies for AI governance. Coordinated international cooperation is essential to address the challenges arising from the rapid development of AI.

Chapter 3: AI Governance for Developing Countries and Advancing Global Sustainable Development

The experts posit that AI governance can provide substantial assistance to developing countries in six ways. AI governance can help developing countries address pressing issues such as poverty, hunger, and health issues by offering precise data analysis and solutions. AI can also help overcome resource constraints by efficiently managing and distributing limited resources, especially in the fields of technology and education. Third, AI can improve digital infrastructure by propelling the development of network and communication technologies, enhancing internet access and computational capabilities, and narrowing the digital divide. In addition, AI governance can reduce capability gaps through providing high quality education and technical training, boosting technical expertise of local talents, and increasing employment opportunities. Fifth, AI can have localized applications that are tailored to local needs and culture, especially when it comes to linguistic and cultural diversity. Lastly, they support international cooperation and support developing countries to participate in international AI governance to ensure that developing countries have a voice in the global AI industrial and value chains.

The experts also make six points regarding the assistance AI can provide for global sustainable development. First, AI contributes by enhancing economic growth through increased productivity, reduced costs, and enhanced global competitiveness, crucial for economic diversification in developing countries. Second, AI also improves social services and infrastructure in areas such as education, healthcare, and urban planning, providing more efficient and precise services, increasing resource efficiency and reducing waste. In addition, AI aids in achieving UN Sustainable Development Goals (SDGs) by monitoring and assessing progress, providing data support for policy making, and managing resources more effectively to reduce environmental impact. Fourth, AI governance can support inclusive growth by tolerantly considering the needs and aspirations of all nations and ensuring technological development benefits the majority of the global population. Fifth, international governance and cooperation is key, so we should create international governance mechanisms and cooperation platforms, promote information and resource sharing, provide economic incentives for compliance with norms, and collectively address global challenges. Lastly, on sensitivity and transparency in AI, we must supervise the practices of the AI industry, to ensure that they fulfill ethical standards, respect data privacy and security, and minimize exploitative practices.

In summary, while AI governance significantly contributes to accelerating development in developing countries and achieving global SDGs, careful consideration of potential challenges and risks is essential, particularly by comprehensively considering different development objectives.

Chapter 4: AI Governance from an Engineering Perspective

Technical methods and tools for supporting governance may include:

- Understanding and assessing model capabilities, which is key. We currently lack a systematic conceptual framework to determine the specific capabilities of models, hindering effective governance of AI. Developing standardized methods for assessing model capabilities should be a priority for governance. The experts also emphasize the importance of forming reasonable strategies and a broad understanding, knowledge, and skills in human-machine interaction. Interpretable AI would also contribute to the development of practical wisdom.

- Strengthening the standardization of AI safety and security governance. This includes establishing mechanisms for updating guidelines, developing specialized standards for application areas, and creating experimental zones. This standardization work can help guide the controllable development of AI.

- Governing generative AI based on risks and multi-party participation, the development of assessment frameworks and tools, and seeking international cooperation. This provides considerations for the responsible application of AI.

- Making AI development and deployment more aligned with public interests through open-sourcing and democratization of governance. Experts suggest building cross-cultural and cross-language ethics databases and remaining open to public participation, which contribute to improving the safety and value alignment of AI systems.

- In addition, it is necessary to strengthen international dialogue and communication, establish inclusive safety and security rules, and open-source high-quality datasets to address current challenges in AI development, such as fragmented rules, difficulties in value alignment, and exacerbated wealth disparity.

Overall, engineering technology plays a crucial supporting role in AI governance. It is essential to deepen our understanding of key issues and translate them into concrete practices in areas such as model design, training, and validation.

Chapter 5: AI Governance from the Perspective of Companies

The discussions from various enterprises focus on different aspects:

- Michael Sellitto introduces Anthropic's Artificial Intelligence Safety Level (ASL) concept, designed to manage the potential catastrophic risks of AI. This approach draws inspiration from the Biosafety Level (BSL) standards for handling dangerous biological materials, defining risk levels based on AI capabilities and requiring different safety measures for each level.

- Jason Si and Jeff Cao from Tencent Research Institute discuss the application of reinforcement learning from human feedback (RLHF) in improving large model alignment, along with the role of other technological and governance measures such as data processing, interpretability, and adversarial testing in ensuring large model alignment. They also discuss methods from the engineering level to ensure the safety and alignment of AI systems

- Wei Tao from Ant Group points out that with the rapid improvements in LLMs over recent years, they have also faced problems including lack of cognitive alignment and issues around principles and interpretability. These issues can lead to serious decision-making errors and rapid diffusion of execution, resulting in unforeseen consequences. Wei Tao suggests that AI systems must enhance cognitive consistency, establish a verifiable chain of reasoning, and engage in interactive learning with human experts.

- Intel's article, using the deepfake detection technology it developed as an example, discusses the responsible application of AI to improve people's lives by enhancing efficiency and creativity, assisting individuals with disabilities, etc. It provides a perspective on responsible AI engineering practices from an applied standpoint.

Overall, these articles, from the perspective of company practices, discuss several noteworthy issues in AI governance, including AI safety tiered classification management, value alignment, open-source governance, and responsible applications. They offer valuable suggestions and examples.

Regarding the Guangdong–Hong Kong–Macao Greater Bay Area (GBA), experts believe that it can provide unique contributions and value to AI governance. Suggestions include:

- Duan Weiwen, Director of the Philosophy of Technology Research Office at the Institute of Philosophy, Chinese Academy of Social Sciences and affiliated with the China Association for Science and Technology - Fudan University Institute of Science and Technology Ethics and the Future of Humanity, proposes three contributions that the GBA can make to AI governance. First, he suggests innovating data governance by exploring the construction of a trustworthy mechanism for data exchange and sharing through institutional innovation and experimentation. Second, he calls for implementing an AI-driven regional integrated development strategy that melds the goals of AI governance with talent, education, and employment strategies in the GBA. This involves strategically positioning talent, education, and industries in the GBA to adapt to the future development of AI. Lastly, he recommends establishing an Eastern Bay Area Special Zone for AI, attracting global talent on a foundation of beneficial AI governance and AI-driven development, and creating a global AI innovation testbed through more dynamic and adaptive industrial promotion policies and continuously optimizing AI governance models.

- Ma Chenghao, Gao Wanqi, and Fan Siyu from the China Electronics Standardization Institute advocate for the establishment of an International AI Alignment and

Governance Innovation Demonstration Zone in Shenzhen. They call on global AI companies and research institutions to participate together, and within a certain scope validate how scientific and how operationalizable are relevant alignment methods, standards and norms, governance tools, data sharing mechanisms, and other such content.

- Wang Jun and Na Diya from the Nancai Compliance Technology Research Institute propose that the GBA should fully explore the value of data, capitalizing upon its advantages in the massive scale of its data and richness of application scenarios. They emphasize the importance of unleashing the potential of data, and on the foundation of data compliance, promoting opening of public data, advancing the construction of multimodal public datasets, and creating high-quality Chinese-language corpora data.
- Nathaniel Sharadin from the University of Hong Kong suggests systematically assessing and understanding model capabilities should be a governance priority. The GBA can leverage its regional advantages to attract international AI companies to come to the area and collaborate in establishing a governance demonstration zone. This zone could validate various AI safety and ethical governance tools, contributing valuable experiences to global governance efforts.

By comprehensively using the unique conditions of the GBA, including its geographical advantages, industrial foundation, and level of openness, the region can actively contribute to global AI governance and enhance its influence on both the region and the nation.

Chapter 1: The Risks and Challenges of Generative AI

The Challenges of LLM Behemoths Toward the Leviathan and Legal Order

JI Weidong (季卫东)

Since Google released the "Transformer" network structure in 2017, in just more than five years, the world has seen the rapid appearance of a huge group of large models, which in turn derive from a variety of technical architectures, a variety of modalities, and a variety of scenarios. In terms of the global distribution of released large models, China and the United States are significantly ahead, exceeding 80% of the global total, with the United States consistently ranking the highest in the world in terms of the number of large models. As soon as ChatGPT was announced at the end of November 2022, it swept the world due to its strong dialogue capabilities and wide range of applications, bringing the monthly active user level to 100 million in just two months, an extremely impressive growth rate. Since then, these large language models (LLMs) have been released one after another, which have profoundly affected various social practice scenarios, including legal operations, from the aspects of empowering individuals and reducing the burden of business, leaving a digital "Cambrian" landscape of the explosion of generative AI species. According to incomplete statistics, by May 2023, Chinese science and technology enterprises and online platforms had launched 79 AI language models of various types, of which 34 were general-purpose models.

It has to be admitted that while the LLMs bring convenience and benefits to the State and society, it also poses disturbing risks and even threats. Four of them can be cited as follows.

- First, as ChatGPT-like large language models provide online dialogue services, they can collect more personal information and privacy than established Internet search engines. Therefore, in the case of "knowing too much and having conflicting interests," LLMs and their operators may induce users to make choices against their own intentions and interests by controlling communication.

- Second, the current stage of the LLM will treat things that do not and cannot exist in the training data as real and describe them in an unquestioning tone in dialogue. This is the phenomenon of “serious nonsense” that users often complain about. From a scientific and technical point of view, this is of course only a kind of “hallucination”. The phenomenon of “hallucination” is closely related to the generalisation ability of machine learning to handle unlimited unknown data with limited training data. However, in application scenarios, this hallucination can lead to the spread of false information, which can be fatal to users or society.

- In addition, LLMs may raise complex issues of intellectual property identification and protection when they use various data for learning or when AI automatically generates various content. In order to ensure the credibility of AI generated content and to clarify responsibilities, digital watermarking techniques should be invented, applied, and promoted.

- Lastly, LLMs may intentionally or unintentionally access confidential information of companies or government agencies, manipulate public opinion, and lead to loopholes in the security system of the central system of the State, dysfunction of the information society, or even unrest due to malicious accidents and crimes.

Much of the human processing of language and harnessing of intelligence actually takes place unconsciously. Philosopher of science Michael Polanyi once noted in 1964, “Man knows far more than he can say.” In other words, the knowledge system should also include such tacit knowledge that is not explicitly realised, or not recognised by the common sense of the society, or cannot be verbalised. This proposition has been expressed as “Polanyi’s paradox” and has become the basis of the theory of artificial intelligence. This also means that AI’s processing of unconscious language simply cannot design the kind of algorithms that acquire and apply all languages, and it is difficult to set clear training goals for machine learning. The now-prevailing machine learning using neural networks gradually reduces the error by continuously adjusting the neuron weights and updating the network parameters through an error back-propagation algorithm to find a positive solution to the training data. It has been found that the accuracy of AI solitaire predictions suddenly and dramatically improves when the size of the neural network is dramatically scaled up. This discovery and its conscious application brought machine learning into the deep learning stage: without the need for complex rules and learning methods, simply multiplying the size of the network can solve many difficult problems and rapidly improve generalisation — it goes without saying that this magical effect also proves the importance of LLMs. In essence, it is the self-learning and in-context learning of multi-layer networks, and the learning of learning methods — meta-learning — that is achieved on this basis. In this way, the design of human features becomes meaningless and the AI actually starts to mould itself and thus forms an automated ecosystem.

It is here that "LLM behemoths" are emerging, and are likely to slip out of human control by abandoning the design of pre-given features in favour of self-imposed sub-goals, raising serious governance issues. This means that LLMs will midwife the birth of a new type of non-human or super-human intelligence that will drift away and develop very different values from those of humans. It also means that in addition to the platform monsters and guerrilla of sovereign individuals hidden in Blockchain, the sovereignty leviathan will face challenges from dozens or even hundreds of powerful large model behemoths, i.e., national sovereignty in the digital realm, or say "digital sovereignty" is facing the challenge of a "War of 100 Models" and loss of control. The concept of "digital sovereignty" clearly reflects the sovereign state's response to the digital transformation of society and its position of self-defence.

In order to prevent the various risks mentioned above from evolving into irreversible disasters, experts and industry leaders have put forward various countermeasures and suggestions, such as suspending the development of large models, achieving value alignment, and strengthening AI regulation. In terms of value alignment alone, for example, the Brookings Institution in the U.S. published an article by Benjamin Larsen titled "The Geopolitics of AI and the Rise of Digital Sovereignty" on December 8, 2022, in which the author argues that the uneven development of AI will lead to growing mistrust between countries, which in turn will lead to the rise of digital sovereignty and the emergence of technological decoupling; the ideological differences or differences in moral principles may have wider geopolitical implications for the management of AI and information technology; thus ensuring the consistency of AI's values at the international level may be one of the most significant challenges of this century. In any case, this is an unprecedented sea change that will inevitably shape new forms of state and legal existence and promote innovation in the paradigm of order.

Given this major change, the Chinese government's strategy is to incorporate and integrate dozens of large model behemoths through a unified basic model, and to prevent the risk of peer-to-peer interaction losing control through the said "sovereign blockchain." The result is bound to create a much more strong algorithm Leviathan. As Michel Foucault had anticipated, this algorithmic Leviathan was actually a Panopticon. Here, billions of probes form a trap of sight, creating the kind of surveillance society and culture depicted by David Lyon. This algorithm Leviathan is ubiquitous and powerful, and its abuse can only be prevented through procedural due process embedded in AI systems and decentralized checks and balances between different AI systems. In this sense, it can also be said that after entering the era of large models and generative artificial intelligence, the focus of AI governance will shift from preventing algorithm discrimination to preventing model abuse. In the interaction of sovereign Leviathan, platform monsters, LLM behemoths and even SSI (Self-Sovereign Identity) consciousness, the principle of legal due process will be redefined and combined with the technical due process, and this kind of new procedural justice will play a more important role.

The challenge of AI value alignment is to build AI systems that align with human values and interests (Russell, 2019). This challenge encompasses technical and normative aspects (Gabriel, 2020). The technical challenge aims to encode human values into AI systems, ensuring they behave as intended. The normative challenge involves determining which values AI systems, and broader AI development efforts, should be aligned to. This article focuses on the normative aspects and explores two forms of AI democratization — the democratization of AI development and the democratization of AI governance — as a means to represent diverse human values in AI development.

Ji Weidong (季卫东), Shanghai Jiaotong University Professor of Humanity and Social Sciences, Director of Center for Japanese Studies (Ministry of Education Accredited Institute), President of China Institute for Socio-Legal Studies, Director of Center for AI Governance and Law, formerly a visiting scholar at Stanford Law School. Recipient of the State Council special government allowance for experts.

Optimizing the Development of Artificial Intelligence through More Proactive Governance

DUAN Weiwen (段伟文)

In recent years, the capabilities of artificial intelligence (AI) in areas including cognition, decision-making, knowledge production, and intelligent agency have increasingly demonstrated extraordinary abilities. This has positioned AI as a top development issue in the technological domain for nations worldwide. However, the prominent dangers, risks, and controversies associated with AI development have prompted each country and the entire world to comprehensively seek to govern artificial intelligence.

The governance of AI primarily addresses two major issues. First, due to the lack of clarity regarding the values and goals of AI development, coupled with malicious use and abuse, AI causes real harm and poses potential risks to humanity, individuals, society, the environment, and ecological systems. This includes concerns such as privacy and data rights violations, exacerbation of bias and discrimination, etc. Secondly, the imbalances in AI development and the unfair distribution of benefits and risks have given rise to societal debates questioning, for instance: “Who benefits? Who bears the costs? Who bears the risks?,” as well as “Who has a leading advantage? Who will be left behind? Who is in an ‘exposed’ state?”

From a risk prevention perspective, urgent global priorities in AI governance encompass three key aspects. First, assessing how to avoid major risks from both intentional and unintentional AI misuse, especially previously overlooked cross-domain, compound risks. This includes unforeseen safety and security risks resulting from unconventional combinations of AI and biotechnology due to increased accessibility of technology and information, as well as lowering of barriers. Second, figuring out how to alleviate the major shock of AI, particularly generative AI, on job openings, employment, talent, and education. Third, exploring how to establish multi-party control mechanisms and open channels of dialogue on the international stage to regulate AI arms races.

Among these issues, the question of how the international community can establish effective risk warning and response mechanisms to ensure humanity has the capability to hit the "stop" button at crucial moments currently lacks a direct answer. However, efforts can be made in four areas. First, each country and the world needs to enhance mutual trust and cooperation regarding the security of information network systems, establishing a global risk warning system. Second, countries, different regions, and the world should establish multiple parallel information network systems and, when it becomes necessary, build a complete data backup "time machine" for human civilization and operate multiple global information network systems in parallel. Third, discover and nurture talents for the AI era who possess exceptional understanding, decision-making abilities, and foresight, and implement education and training plans for extraordinary talents in each country. Fourth, strengthen sociological, anthropological, psychological, and philosophical research on human-machine behavior, and conduct systematic exploration of this issue.

From the perspective of promoting development, in order to assist all countries, particularly developing countries, in high-quality development and achieving the United Nations Sustainable Development Goals, AI governance should start from two aspects: "passive compensation" and "proactive improvement." The first involves seeking to alleviate the aforementioned issues to foster responsible and trustworthy development of AI systems throughout their entire lifecycle. The second involves the international community jointly taking a series of additional proactive measures, including implementing compensation and balanced development policies, anticipating risks, strengthening prevention, and enhancing AI literacy in underdeveloped regions.

In conclusion, for the Guangdong–Hong Kong–Macao Greater Bay Area (GBA) to participate in and contribute to AI governance, I have three recommendations. First, innovate in data governance by exploring trustworthy mechanisms for data exchange and sharing through institutional innovation and experimental exploration. Second, implement an AI-driven regional integration development strategy, melding AI governance goals with talent, education, and employment strategies, structuring the layout of talent, education, and industry in the Greater Bay Area, adapting it to the future development of AI. Lastly, establish an Eastern Bay Area Special Zone for AI, attracting global talent on the basis of beneficial AI governance and AI-driven development, and construct a global AI innovation experimental zone through dynamic and adaptive industrial promotion policies and continuously optimizing AI governance models.

Chapter 1: The Risks and Challenges of Generative AI

Duan Weiwen (段伟文), Professor of the Department of Philosophy of Science and Technology in the Institute of Philosophy, Chinese Academy of Social Sciences (CASS) and director of the Research Center for Science, Technology and Society at CASS. Professor of the China Association for Science and Technology - Fudan University Science and Technology Ethics and Future of Humanity Research Institute, recipient of the State Council special government allowance for experts.

Challenges in Aligning Human Values and Embedding Ethics in Large Models

WANG Xiaohong (王小红)

Information ethicists assert that it is necessary to establish a fundamental set of ethical principles for AI, but this is a difficult task due to divergences of ethical principles across various cultural contexts and AI usage scenarios (Taddeo and Floridi, 2018). Machine ethics philosophers emphasize that while there exist values shared by humanity that transcend cultural differences, there remain nuanced differences across cultures and human moral frameworks (Wallach and Allen, 2017:66). Empirical research suggests that the effectiveness of AI ethical principles lies in their localization, necessitating adherence to local cultural, religious, and philosophical traditions during the localization process (Danit Gal 2019:73). These studies highlight that in practical AI governance scenarios the abstract principle of "human-centric" often will result in divergent practical values due to cultural differences, potentially resulting in mutual counteraction of AI governance techniques.

Therefore, given the difficulty and complexity of aligning with human values, we propose a consensus strategy for construction of AI ethics informed by philosophical wisdom:

First, transcend the cultural "conceptual pigeonhole" (Dewey, 1921: 188). In dialogue between diverse cultures and values, individuals often employ their familiar cultural "conceptual pigeonhole," compartmentalizing facts from another culture to interpret different cultural phenomena. This leads to hasty categorization and subjective judgments. Understanding the true intentions of a different culture is always a complex task. Different cultures carry the unique and extensive life histories of different human communities, and “the actual reason for making history is the will to survive and the desire for happiness. But what is happiness? Answers to this question are very different. This is because we have many different philosophical systems, many different value standards, leading to many different types of history" (Fung Yu-lan, 1922). Humanity can only correct itself through self-reflection, relying on reason.

Second, under principles conducive to human progress, respect the diverse values baked in by different histories. Comparative studies of Chinese and Western philosophy reveal that an important aspect of cultural differences lies in the fact that while all cultures share some basic values, different cultures may assign different weights to these values, forming distinct configurations of values (Li Chenyang, 2019). Different cultures may never converge on value configurations, but diverse cultures can achieve consensus favorable to human progress and development of humanity on the basis of a diversified configuration.

The philosophical thoughts and ways of thinking in the East and the West, spanning thousands of years, not only remain timeless but also exhibit profound resonances since ancient times. Confucius from pre-Qin China stated: "Do I possess knowledge? I have none. A rude fellow questioned me, and I responded with emptiness. I knocked at its two ends and found it empty" [吾有知乎哉？无知也。有鄙夫问于我，空空如也。我叩其两端而竭焉]. In ancient Greece, Socrates asserted, "All I know is that I know nothing." These two philosophers from the East and West, almost in the same era (referred to as the Axial Age by Karl Jaspers), expressed resonant maxims for scholarly endeavors. Several expressions in the Analects are consistent with the key ideas of Kant's moral law, i.e., only when you are willing to act according to this principle does it become a universal norm. Originating from the concept of "cautious solitude" [慎独] from the "Book of Rites: Doctrine of the Mean" 《礼记·中庸》, the Neo-Confucians developed the concept of "cultivating oneself to bring about social order" [修齐治平], and these ideas align with Aristotle's advocacy of virtue ethics, reflecting the shared philosophical wisdom of ancient cultures in the East and the West.

Third, embedding ethics in large AI models, whether from top-down or bottom-up, requires a thorough analysis and clear presentation of the semantic implications of moral philosophy. The current successful training paradigm for large models (LMs), based on "high-quality dataset construction—large-scale pre-training—instruction fine-tuning—reinforcement learning from human feedback," integrates symbolic (top-down) and connectionist (bottom-up) approaches, and has been referred to by computer scientists as the future priority development direction of "integrated intelligence" (Chen Xiaoping, 2020:116). In this regard, contextual analysis and evaluation based on moral semantics are required when constructing reasoning-based knowledge bases or search-based state spaces, as well as training large models with neural networks incorporating ethical rules and utilizing representative databases that involve implicit ethical principles. To obtain consensus on human values, we propose developing computational hermeneutics similar to modeling of the Han Dian《汉典》Ancient Classics

(Wang Xiaohong et al., 2023), enabling the clarification and differentiation of the rich meanings of moral concepts at the formal level and integrating cultural significance into supervised (manually annotated) and unsupervised (automatically annotated) machine learning.

Wang Xiaohong (王小红), Professor of Philosophy in the Department of Philosophy at Xi'an Jiaotong University, Co-Director of Computational Philosophy Lab. Researches philosophy of cognitive science, AI machine discovery, analysis of meaning (Handian LDA-TM), AI & big data ethics, science & humanities (HPS).

The Global Regulation of AI: Major Gaps and Principal Challenges

Rostam J. Neuwirth

New technologies, commonly referred to as "artificial intelligence (AI)", are being introduced at a faster pace and increasingly pervade more and more aspects of human life. As reflected in the adoption of the UNESCO Recommendation on the Ethics of AI in November 2021, the initial hype about the potential benefits of AI has now been supplanted by ethical concerns about its actual and potential harms. In attempts to address these concerns, the world is witnessing a global race to regulate AI through binding legal instruments. The European Union, the Council of Europe, the People's Republic of China, the United States and many more jurisdictions have already adopted or are in the process of preparing laws specifically or generally targeting AI.

The present global AI race, however, poses serious challenges to law and the existing international legal framework, which include, first and foremost, a strong temporal element, which means to not only find the optimal moment in time to enact AI laws or regulations, but also to future-proof them to secure their adequacy for a meaningful time in the future.

Second, it also features a spatial dimension, which consists in legal problems caused by the ubiquitous or cross-boundary nature of AI, which stands in a stark contrast with the traditional territorial conception of laws. To account for the interoperability of various AI systems and to prevent possible technical breakdowns or norm conflicts, a multi-level governance approach must be adopted, which enables the global coordination and harmonization of various local, national or regional regulatory approaches.

A third challenge is found in the all-pervasive effects of the impact of AI on societies, organisations and humans. It means that AI is a cross-cutting phenomenon, which requires first an interdisciplinary debate to support the formulation of a coherent regulatory approach based on the establishment of a consistent institutional framework allowing for a more efficient form of multi-agency cooperation.

A fourth challenge lies in the limitations of present languages to better embrace the novel characteristics of AI and related technologies. This problem emerges in the qualification of AI as an oxymoron, i.e. a figure of speech, which links intelligence to the two apparently contradictory or highly dissimilar terms of humans and machines. The ensuing contradiction, hence, requires us to rethink the fundamental premises of human cognition and to clarify the principal objectives of AI regulation, namely whether it aims to focus on the technology, the businesses or providers putting it into the market or the persons using them. In this regard, current proposals are often too vague, or both too specific and too general at the same time. For this reason, a truly comprehensive regulation of AI is in need of novel modes of thinking that simultaneously match specific legal questions against the consistency of the legal system as a whole. More importantly still, the complexity of all these challenges combined warrants a global philosophical consensus about both the potential usages and wider purposes of AI for the evolution of humanity.

Rostam J. Neuwirth, Professor and Head for the Department of Global Legal Studies at the University of Macau Faculty of Law. Research interests include International Economic Law / International Trade Law / WTO Law, European Union (EU) Law, and Transnational Law.

Responding to the Challenges of Generative AI on Global Governance

SUN Nanxiang (孙南翔)

At present, the rapid advancement of emerging technologies, particularly generative artificial intelligence (AI), is ushering in a new era of technological and industrial revolution. The application of AI technologies not only catalyzes the transformation of national governance systems but also profoundly influences the development trajectory of global governance mechanisms. Generative AI technology introduces a capacity for "self-awareness and autonomy" in technological entities, endowing non-state actors with power comparable to that of nation-states. In light of these transformative dynamics, the role of the law must be reassessed and reimagined, especially in ensuring that technological development aligns with the trajectory of human development. This imperative extends to the realm of global governance.

Compared to the traditional era, the age of artificial intelligence will overturn traditional international legal frameworks. Key elements of the international system, including its actors, structure, and operating rules, will undergo substantial changes. The development of AI technology poses formidable challenges to the global order but also presents new development opportunities. Currently, generative AI tools are widely employed globally in media, research, and even transportation, armed conflicts, and various other domains. It is critical to determine the legal properties of AI tools, however as of now human society has not yet come to a consensus. In the foreseeable future, the self-learning and self-awareness capabilities produced by generative AI technology cannot be readily regulated or predicted by humans. Undoubtedly, the world's ignorance and lack of knowledge about AI technology is a significant challenge.

Examining the current stage of AI technological development, generative AI technology originates from and is subject to human control. The thoughts, concepts, and cognition generated by AI are derived from the human world. We should accelerate research on the impact of generative AI technology on domestic and international legal mechanisms and AI's challenges to human society and life. Overall, we should adhere to the principles of systemic integration, technological permeation, and legal technologization when addressing the challenges posed by AI to international law. From this perspective, nations should actively explore the constraining role of rule of law principles on AI technology, strengthen the integration of morality, ethics, and technology, and collaborate with nations around the world to collectively confront the challenges triggered by generative AI technology.

Sun Nanxiang (孙南翔), Chinese Academy of Social Sciences International Law Research Institute International Trade Law Research Office Associate Researcher, primarily researching international economic law and internet law. Member of the board of the China Law Society Cyber and Information Law Society.

Global AI Risks Inescapably Require Global Collaboration

Duncan Cass-Beggs

Artificial super intelligence could be closer than we think. This raises urgent existential questions for humanity such as whether and when it is safe to develop such advanced AI systems, and what role such systems should play in society. To be effective and legitimate, these choices need to be made – and implemented – collectively by the global community. This will require unprecedented global collaboration within potentially short timelines and must succeed despite a context of geopolitical tension and conflict. Steadfast resolve and tireless innovation will be needed to meet this challenge.

Global-scale risks from AI can no longer be ignored. The existence of these risks rests on three observations: 1) AI systems are developing rapidly and could far surpass human-level performance across a wide range of crucial capabilities. This could occur sooner than expected. 2) Humanity currently lacks the means to reliably control such systems or otherwise ensure that they remain aligned with human interests. Some experts argue that it may ultimately prove impossible for a lesser intelligence to reliably and sustainably control a vastly superior intelligence on an ongoing basis. 3) If humans create extremely capable AI systems that cannot be reliably controlled, the outcome is most likely to be harmful. Examples of catastrophic global scale risks from advance AI include a) misuse by malicious actors, such as the creation and widespread deployment of novel pathogens or cyberweapons, and b) misalignment, which refers to the creation of one or more powerful autonomous AI systems with goals that conflict with those of humanity and that cannot be controlled or stopped.

The first and most urgent question that requires globally coordinated decision-making and enforcement, is about whether and when it is safe enough to allow the development of superintelligent AI systems. People in all parts of the world share a common interest in ensuring that nobody, anywhere, develops an AI system that could potentially imperil humanity. AI systems with super-human general capabilities should therefore not be permitted until reliable alignment and control mechanisms are in place. To achieve this goal, a coordinated licensing regime could require prior risk assessments and permission to develop the most powerful (i.e. “frontier”) AI systems, and such a regime could be backed up by robust international monitoring mechanisms. Enforcing this regime would become more challenging, however, as the algorithms, data and computing power needed to develop potentially dangerous powerful AI systems become more widely accessible.

The second question that humanity must address collectively is what future to pursue if ever it is possible to develop safe and aligned artificial superintelligence. Guaranteeing safety may take a long time or even prove impossible, requiring a prolonged or perpetual water-tight global prohibition on the development of artificial superintelligence. However, if at some point humanity determines that it is safe to create one or more superintelligent systems, many fundamental questions arise such as how many AI systems to create, and what goals, rights or limits such systems should be given. Unlike many AI policy choices that can be made independently by different countries, questions around the introduction of a new highly capable species of intelligence would need to be made collectively because such AI systems would have impacts, at least indirectly, even on societies that might not wish to host such systems. In theory, humanity could seek advice on some such questions from an AI itself, but at a minimum some agreement will be required in advance on the parameters and framing of these questions and the kind of AI(s) selected to ask them to.

The transformational potential of AI, long the preserve of science fiction, is now realistically at humanity’s doorstep, requiring key decisions and action in the months and years ahead. Despite the magnitude of these issues, it seems far from assured that humanity will recognize them or rise to the challenge in time. Hard work will be needed to help decision-makers understand the risks and to develop the new paradigms and institutions required to navigate this period successfully. Imagination and commitment will be needed from all parts of society and all parts of the world to make this happen.

Duncan Cass-Beggs, Executive Director of the Global AI Risk Initiative at the Centre for International Governance Innovation, former Counsellor for Strategic Foresight at the OECD.

Chapter 2: Global Governance Approaches for Generative AI

Institutional Design Principles for Global AI Governance in the Age of Foundation Models and Generative AI

Nicolas Moës, Yolanda Lannquist, Niki Iliadis, and Nicolas Miailhe

The emerging era of foundation models and generative AI poses new risks and challenges for AI governance. Given their wide application and rapid adoption, it's crucial to understand and mitigate these risks proactively. This essay proposes ten institutional design principles for AI governance to ensure safety and uphold human values.

1. **Intrinsic, broad and unpredictable risks:** Risks are use case and application-agnostic, and arise across the lifecycle, from design to deployment. Failures in these systems could pose large-scale, catastrophic or even existential risks emerging in unpredictable ways, from widespread biased content moderation and public mental health crises to automating malicious use, biosecurity, cybersecurity, and national security threats.

2. **Trustworthiness-by-design**: This approach aims to embed safety, security, and ethical considerations into the AI from the onset, rather than retrofitting them later, enabling creation of new markets relying on trustworthy models.

3. **Take back control:** Third-party evaluations by independent experts should supplement internal assessments to provide a robust safety net. Regulatory authorities must have the power to halt or amend the development process based on these evaluations. It is critical to ensure a robust institutional framework beyond private sector-led internal assessments.

4. **Tech- and channel-neutral:** To ensure a level playing field and mitigate unforeseen risks, regulation should be applied across technology and distribution channels. Whether an AI system is distributed through open-source platforms or APIs, or developed using varying technological paradigms like brain emulation or rule-based systems, the baseline level of regulatory scrutiny should remain constant.

5. **Targeted liability and responsibility:** Powerful stakeholders like big tech companies and artificial general intelligence (AGI) companies should be responsible and accountable for impacts of their products, as the primary bearers of liability. Within these entities, roles like safety officers and compliance officers should also have specific obligations.

6. **Structural and systemic practices**: Practices for safety, ethics and security should be embedded in an organization's culture, rather than confined to ad-hoc, add-on measures like red teaming. Regulations should enforce evidence-based requirements that evolve with the state of the art in AI capabilities.

7. **Evidence-based requirements:** Developers should be obliged to demonstrate the effectiveness of their safety, ethics and security practices through empirical evidence. Evidence-based measures should evolve with the state-of-the-art AI capabilities and risk mitigation and prevention measures, ensuring future-proof systems.

8. **Public sector capacity-building:** Enhance regulatory effectiveness, which relies heavily on the knowledge and skills within the public sector. This knowledge should be acquired independently and not be unduly influenced by industry. A well-informed public sector can better resist regulatory capture and make informed decisions on AI governance amidst technological evolution.

9. **Adaptable and resilient governance mechanisms:** Policy continuity must be flexible for updating based on evolving circumstances while avoiding diluting or shifting scope due to industry capture. New institutions, such as a European Union AI Office, should be empowered to update rules. Oversight from civil society can ensure that focus remains on safety, ethics, and security.

10. **Interoperable global governance:** Promote coherent AI regulations across jurisdictions that are mutually reinforcing in implementing the most robust safety,

ethics, and security measures. Entities should coordinate to "raise the bar" in terms of governance requirements and avoid fragmentation which can lead to regulatory arbitrage.

These principles offer a multi-faceted, evolving approach to governance that can adapt to technological advancements. Abiding by them can help ensure AI risks are mitigated and benefits are distributed widely and fairly.

Nicolas Moes, Director for European AI Governance at The Future Society (TFS), appointed expert at the OECD.AI Policy Observatory, and focusing on European developments in the legislative frameworks surrounding AI, including the EU AI Act drafting and enforcement mechanisms.

Yolanda Lannquist, Director of Global AI Governance at The Future Society (TFS), appointed expert to the OECD AI Policy Observatory, leading TFS's AI governance & policy projects with international organizations, governments, companies, academia and nonprofits for AI safety, ethics, security and inclusion.

Niki Iliadis, Director of AI and the Rule of Law at The Future Society (TFS), leading the international and multistakeholder forum The Athens Roundtable on AI and the Rule of Law and U.S. AI policy.

Nicolas Miailhe, President and Co-founder of The Future Society (TFS), an appointed expert to the Global Partnership on AI (GPAI), OECD's AI Group of experts (ONE.AI), and UNESCO's High Level Expert Group on AI Ethics.

Key Policy Recommendations for Global AI Governance

ZHOU Hui (周辉)

In an era where artificial intelligence (AI) is continuously reshaping economic and social structures, establishing wise governance mechanisms for this transformative technology has become exceptionally urgent. Examining China's "Artificial Intelligence Law Model Law" (hereafter, “Model Law”) and explorations in governance by the European Union and the United States, reveals the need for a multidimensional framework to effectively supervise the research and development (R&D) and deployment of artificial general intelligence (AGI) or foundation models beyond a certain scale. This framework should establish effective risk warning and response mechanisms and, adapting to the development patterns of AI, promptly address new risks and challenges arising from it.

1. Formalization of Ethical Principles

The "Model Law" explicitly requires the use of system architectures that allow human supervision and intervention for AI R&D, provision, and use. This system design aims to ensure the ability of humans to correct or stop the operation of AI systems, preventing harm from loss of control of the technology. In addition, for better early warning of risks, the “Model Law” incorporates a series of internationally recognized AI ethics principles, such as ensuring transparency, fairness, and accountability. Researchers, developers, and providers of AI need to not only conduct labeling for improved user awareness but also should ensure transparency and interpretability of AI algorithms and models in design, and they should provide relevant information or release a statement to the public or regulatory authorities when necessary. In terms of fairness, the “Model Law” requires that AI R&D and application promote inclusivity to reduce the digital divide and serve various marginalized groups. It also requires AI systems to minimize the generation and output of discriminatory content as much as possible and restrain users from engaging in similar activities.

2. Precise Governance

Given the inherent differences in complexity and societal impact of AI technologies and variations among AI systems of different scales and uses, it is difficult for a single governance framework to meet the needs of precise governance. The "Model Law" proposes a governance model based on a negative list, using post-hoc supervision and reviews as the cornerstone, while incorporating preventive strategies for high-risk AI R&D and provision. This hybrid approach echoes the risk-based framework explicitly articulated in the European Union's "AI Act." For AGI or foundation models, the “Model Law” also imposes specific obligations on R&D entities and providers, drawing from mature practices and examples in platform governance for “digital gatekeepers” to fulfill special responsibilities and obligations, in order to ease regulatory pressures and improve governance efficiency.

3. Balancing Development and Security

As AI technologies assist in economic and social development, governance measures need to strike a balance between development and security, avoiding overly strict institutional designs that hinder technological innovation. Therefore, regulatory agencies should create a more certain regulatory environment to ensure the business prospects of relevant enterprises. Additionally, the “Model Law” incorporates several systems explicitly aimed at promoting the development of AI technology, including developing computing infrastructure construction and other supporting areas, facilitating supply of data factors, designing regulatory sandbox systems, and exempting open-source AI technology providers from certain responsibilities. Furthermore, the negative list system also aims to create a more permissive environment for the development of low-risk AI technologies. Apart from the AI activities listed in the negative list which require administrative permits, other AI R&D and provision activities only need to complete filing procedures related to information disclosure, which should not become administrative obstacles and require substantive review and approval. The existing legal framework can still be used for low-risk AI activities.

4. Encouraging Technological Governance and Pluralistic Governance

The "Model Law" encourages the development of regulatory technology and compliance technology, implementing a vision of comprehensive, multi-stakeholder governance of AI. On the one hand, support should be provided for the R&D and application of regulatory technology and compliance technology, innovating governance tools, reducing governance pressure, and promoting the intelligence, automation, scientific nature, and precision of governance. On the other hand, AI companies play a crucial role in governance. They have the conditions to integrate values and legal rules into AI systems and have sufficient motivation to participate in AI governance to create a stable development environment for themselves. It is important to establish effective communication and consultation mechanisms, advocate for company self-discipline and industry self-governance, and leverage the role of public opinion in supervision, to improve the comprehensive governance system.

5. Establishment of a Dedicated AI Governance Authority

In the current global governance of AI, given the transregional, cross-domain, and transnational applications that AI can achieve in conjunction with existing network information technologies, addressing potential issues of duplicated and redundant regulation becomes urgent. The "Model Law" advocates for the establishment of a sole Chinese national regulatory authority on AI, which would assume the responsibilities of AI regulation while coordinating with other departments. This institutional design aims to enhance the consistency, professionalism, and certainty of governance by consolidating regulatory functions under a dedicated regulatory body. Specialized AI regulatory authorities in each country would also be able to efficiently establish and participate in international dialogue and coordination mechanisms, promoting cooperation such as the interlocking of AI governance rules between countries and regions, as well as mutual recognition of rules.

Zhou Hui (周辉), Associate Professor and the Deputy Director of the faculty of Cyber and Information Law at the Institute of Law, Chinese Academy of Social Sciences, primarily studying data governance and the regulation of artificial intelligence (AI). Visiting Scholar at Yale Law School.

International Oversight for the Development and Deployment of Foundation Models

Robert Trager and Fynn Heider

Establishing international oversight of a powerful emerging technology is challenging. Yet it is important that we do so for advanced artificial intelligence (AI). Risks from the technology cross borders. Even as some states lead AI development, all are affected, and thus all should have a voice in its development and deployment regulation.

In this piece, we focus on opportunities for international governance of civilian AI. Military AI governance is important but even more challenging. We can see this from the experience of trying to regulate Lethal Autonomous Weapons through the UN Convention on Certain Conventional Weapons. Despite ten years of effort, no legally binding measures have been adopted.

Prospects for international regulation of civilian AI are brighter. One approach would be to use a model similar to the International Civil Aviation Organization (ICAO), International Maritime Organization (IMO), or Financial Action Task Force (FATF). These organizations do not audit firms; they audit jurisdictions. An International AI Organization (IAIO) could audit jurisdictions to make sure they adopt international regulatory standards and have a track record of enforcing them. Like the IMO and ICAO, such an organization should work closely with states to ensure they have the capacity for compliance.

Every governance regime needs to incentivise compliance. One approach to incentivizing compliance in a civilian AI governance regime is to tie it to trade. States might agree not to import goods whose supply chains include AI from jurisdictions in significant violation of IAIO standards. They might also refuse to export key inputs to such jurisdictions or offer technological assistance to states in compliance with the agreed frameworks.

Resulting monitoring and enforcement at the domestic level could bring major benefits. States will be less concerned about proliferation, and local monitoring and enforcement may facilitate rapid, effective correction of violations. Domestic law enforcement, as a rule, moves faster than international equivalents.

This framework leaves open what the specifics of foundation model regulations should be — we have generally not yet reached consensus on those, domestically or internationally. But neither is full consensus needed to begin setting up international governance; only broad agreement on a minimal set of standards is.

International actors might decide that foundation models over a certain size would be subject to regulatory scrutiny including a standardized set of model evaluations. Development and deployment of such systems could be licensed by local authorities. Data center operation would also need a license that requires "know your customer" practices similar to the financial services industry.

Such a regime could limit harmful development of AI while also facilitating broad access. In fact, a regulatory regime may be necessary for broad access. Without it, frontier states are likely to worry about proliferation and restrict access to the technology more tightly. A global regime could encourage participation from frontier AI states while still guaranteeing the right of voice to affected communities across the globe.

Robert F. Trager, Director of the Oxford Martin AI Governance Initiative, International Governance Lead at the Centre for the Governance of AI, and Senior Research Fellow at the Blavatnik School of Government at the University of Oxford, recognized expert in the international governance of emerging technologies, diplomatic practice, institutional design, and technology regulation.

Fynn Heide, Research Scholar at the Centre for the Governance of AI. He studies international AI safety collaboration as well as AI safety and policy in the PRC.

Coordination, Cooperation, Urgency: Priorities for International AI Governance

Carlos Ignacio Gutierrez

The international governance of artificial intelligence (AI) is an inherently complex problem for which, as of late 2023, we have no clear solution. As this technology's capabilities increase at an unpredictable rate, over 190 national jurisdictions are tasked with managing its escalating and unforeseeable risks. To address these risks, independent action is clearly a sub-optimal approach since patching a problem in one place, will not prevent its spread to others. Instead, the effective governance of AI is a communal effort that requires global participation. Considering this, society should prioritize the development of a multilateral response that considers the following: What elements of AI should be governed and how? Who should be included in this governance process? When is the right time to act?

What and how: The options space for AI risks is characterized by its breadth and depth. Any number of issues could be proposed to pool the international community's attention. However, to succeed in catalyzing action, a candidate issue has to trigger a sense of commonality among countries with wide-ranging needs and capabilities. A proposal to jump-start the conversation is to focus on the mitigation of shared large-scale high-risk harms caused directly or indirectly by AI systems. The benefit of setting such a threshold is that it encompass a relatively narrow set of concerns. Moreover, it serves to centralize awareness and synchronize efforts, optimally through a multilateral organization with a concrete workstream composed of the following objectives: Identify vectors of shared large-scale high-risk harms produced by AI systems. Although many concerns will emanate from general purpose AI systems, the effort's remit must include narrow systems that qualify under its operating guidelines. Moreover, it should proactively inform stakeholders on potential risks and recognize existing vectors of harm. Coordinate global responses that are technically sound and consistent with best governance practices. This can take several shapes and depends on the scale and source of the problem at hand. For instance, a response can range from the solicitation of voluntary standards as a precautionary measure to the imposition of a compulsory set of rules as a reaction to an ongoing concern. Enforce adherence to agreed-upon actions that reduce the likelihood and impact of harms. Because the direct and indirect effects of AI are often unbound by jurisdiction, establishing an effective enforcement regimen requires maximizing the number of participating states. While the multilateral effort should be empowered to perform this role, it may also recruit, certify, or deputize public institutions and third-parties, often on a jurisdictional basis, to take on this task in order to scale its enforcement capabilities.

Who: Regardless of their capability to design, develop, or deploy AI technologies, all countries are vulnerable to AI's risks, and may wittingly or unwittingly host or shelter any part of the high-risk AI supply chain or the system itself. This is why cooperation must become a priority regardless of geography, a country's political system, or ideology. Essentially, no state should be excluded from engaging in multilateral action to address global AI concerns. The ability to mitigate shared large-scale high-risk harms depends on wide-spread participation. Thus, incentives should be considered for a range of countries, from those that are influential in the design, development, and deployment of AI to those with a role relatively limited to being subject to this technology‘s risks.

When: We face conditions where evermore powerful AI systems are deployed on a daily basis, and limited bandwidth is devoted to understanding what constitutes appropriate governance. This underscores the urgency of establishing a collective effort to proactively address AI risks. Even if efforts are undertaken today to begin multilateral coordination and cooperation, years will likely pass before a system is put in place. Therefore, in the short-term, it is understandable if an influential initial set of countries take the initiative to begin this multilateral process. This may include the membership of states that lead the world in the commercial deployment of systems, manufacturing of hardware, educating the technology's workforce, and/or establishing comprehensive regulation. In the long-term, the UN is the only organization with universal representation and the ability to host an effort such as the one described in this commentary. Ideally, it takes the reigns over the verification, coordination, and enforcement of efforts to mitigate the shared AI risks.

Conclusion: In optimizing multilateral governance, no "right" answers exist. What we can hope for is a multilateral AI governance scheme that prioritizes coordination, cooperation, and urgency in addressing shared large-scale high-risk harms. By focusing global attention on the mitigation of these issues, the international community needs to build the necessary commonality to achieve the only responsible end state for AI governance: one where the design, development, and deployment of this technology is safe and ethical.

Carlos Ignacio Gutierrez, artificial intelligence (AI) policy researcher at the Future of Life Institute, focusing on the impact of this technology's methods and application on hard law and AI's management through the design of effective and credible soft law programs.

Data Ethics and the UNESCO Recommendation on Open Science

Committee on Data of the International Science Council (CODATA) Data Ethics Working Group

The growing application of big data and artificial intelligence (AI) in scientific research raises ethical and normative challenges, particularly in relation to openness, privacy, transparency, accountability, equity, and responsibility. The Data Ethics Working Group （DEWG） of CODATA is working with global scholars to collaboratively establish a basic consensus for further activities and research on data ethics principles and a data ethics framework covering the whole data life cycle. DEWG has 4 thematic groups which focus on Scientific Integrity, Protection of Personal Data, Indigenous Data Governance, and Global Power and Economic Relations. This will help CODATA to advance its mission in championing global open data exchange and applications in alignment with the UNESCO Recommendation on Open Science.

The Ethics and Scientific Integrity thematic group discussed the topics of transparency, quality, reusability and impact of research as well as management and interpretation of research data, with a focus on collaborative efforts and the role of open scholarship in supporting research integrity. We recommended to establish policies and practical guidelines to advance global norms on data ethics, and foster data sovereignty of researchers by creating support structures, and strengthen the role of research methods as a key part of data ethics, and develop training and educational resources on data ethics.

The Ethics and Protection of Personal Data thematic group explored questions including the politics and political economy of data, and we recommended to build a framework and policy based on a more critical understanding of privacy that recognises underlying dynamics of harm and power that impact individuals and communities, and provide opportunities to upskill in technical, social, legal, and political developments concerning privacy and related topics.

The Ethics and Indigenous Data Governance thematic group focused on data principles such as CARE (Collective benefit, Authority to control, Responsibility, and Ethics) and JUST (Judicious, Unbiased, Safe, and Transparent), and we recommended that Indigenous data sovereignty needs institution building for data trustees (and similar intermediaries) which would enable selective digital disclosure.

The Global Power and Economic Relations thematic group explored the structural conditions shaping research at a national and individual level . Scholars in many national contexts face barriers such as lack of basic infrastructure, unsupportive national policy, the control of the research agenda by Global North funders, and the domination of oligopolistic publishers and Big Tech companies. At an individual level, researchers everywhere who do not fit the expected norm of a scholar (white, able-bodied, male) face multiple barriers such as conscious and unconscious bias, racism, misogyny, career breaks, and societal expectations about caring responsibilities.

DEWG has been approved for promotion to the Data Ethics Task Group at the CODATA General Assembly 2023 which was held on 27-28 October in Salzburg, Austria. Facing the challenges caused by the rapid development and widespread application of artificial intelligence, this task group will strengthen research on data access ethics in our subsequent work.

Acknowledgements: This document is based on the research results of DEWG thematic groups and reports which were released in International Data Week 2023 by Johannes John-Langba, Joy Jang, Masanori Arita, Louise Bezuidenhout and Lianglin Hu.

CODATA: As the Committee on Data of the International Science Council (ISC), CODATA helps realise ISC's vision of advancing science as a global public good. CODATA does this by promoting international collaboration to advance Open Science and to improve the availability and usability of data for all areas of research.

International Governance of Artificial General Intelligence and Large Foundation Models: Progressive Principles and Open Exploration

ZHANG Peng (张鹏)

The deployment and application of artificial general intelligence (AGI) or large foundation models can significantly enhance productivity and improve decision quality, and it is a revolutionary technology that can influence the development of human society. However, in practice, these systems may also introduce issues involving bias, discrimination, data privacy, network security, intellectual property, and misinformation. Without the implementation of safety and security safeguards, these technologies could adversely affect human life and even pose existential risks.

The main difference between AGI or large foundation models compared to normal AI is the massive scale of their data and parameters, and their operating processes are difficult to interpret. It is a resource-intensive and data-intensive industry. The large-scale commercial applications of these models form a supply chain composed of model providers, deployers, users, and consumers. Consequently, the governance or regulation of AGI or large foundation models should focus on individual entities, nor should it take an approach of "joint responsibility" and "collective accountability" for all participants. Instead, governance and regulation should employ systemic thinking and a precise understanding of the supply chain, distinguishing between different entities along the supply chain and their control over the system. Based on the degree of control, detailed and differentiated safety and security obligations and compliance requirements should be established. For instance, upstream providers should bear more safety and security responsibilities for the risks associated with underlying models, while downstream users should assume additional obligations for preventing, mitigating, and eliminating safety and security risks arising from application of foundational models by their products and services.

The international governance of AGI or large foundation models would ideally involve the development of unified rules within the most universal and representative framework under the United Nations. Some viewpoints suggest drawing lessons from the international community's governance experiences with nuclear energy and climate change, aiming to establish similar systems and institutions. However, it should be acknowledged that AGI or large foundation models are an emerging technology and still are in nascent stages of development, so they touch on national sovereignty and complex factors such as geopolitical dynamics and technological competition. Various countries may have differing positions and views of their national interests, making it challenging to achieve consensus and establish clear rules. Therefore, it is important to adhere to fundamental principles of simultaneously developing, exploring, and governing, prioritizing ethics and standards, and step-by-step advancing careful and inclusive domestic and international legislation. Effective international governance frameworks should be established while simultaneously ensuring technological advancement and innovation, and eliminating the potential risks associated with AGI or large foundation models.

Zhang Peng (张鹏), Senior Researcher at the Shanghai Association for AI and Social Development, researcher at the University of International Business and Economics Digital Economy and Legal Innovation Research Center.

Chapter 3: AI Governance for Developing Countries and Advancing Global Sustainable Development

How far are we from harnessing the power of AI in developing countries?

Eugenio Vargas Garcia

For many developing countries, the promise of artificial intelligence (AI) still looks a distant dream. Global inequality is a tough nut to crack. The magnitude of the challenge should not be underestimated. The reasons for this are manifold. Here are four of them:

- Competing priorities – More urgent problems, such as poverty, hunger, health emergencies, or violence, usually take precedence in policy decisions and budget allocations.

- Lack of resources – Trying to solve different issues at the same time with little funding to support it may prove an intractable task, especially regarding science and technology.

- Poor digital infrastructure – Inefficent connectivity (at times unreliable electricity too), outdated communication networks, shortage of hardware and computing power, and so on.

- Capacity gap – Although talent may be found everywhere, regardless of nationality, when quality education, technical expertise, and job opportunities are missing, the effort required to change this is much harder.

Due to the above reasons and other underlying factors, numerous Global South countries have been struggling to bridge the computing divide and become AI-ready nations. Power asymmetries, widening wealth disparities, data exploitation, cyber-colonization, underpaid annotators, algorithmic bias, and discriminating AI systems are among the most common predicaments.

In many places, technology remains underdeveloped and underused. End-users have online access to products and services developed elsewhere, possibly trained with biased datasets that may not be suitable for local needs or national priorities.

More models should be trained in less-known native languages in order to reach out to as many people as possible, including disadvantaged groups or vulnerable communities. Technologies are rarely neutral. Generative AI has been dominated by software from (and for) English-speaking users. Developing countries must promote research on AI that speaks their language.

Much has to be done to ensure that AI can help humanity end poverty, build resilient societies, protect the planet, and achieve the Sustainable Development Goals by 2030.

Trying to catch up through large amounts of investments is not for everyone, in particular least developed countries facing serious hurdles. Full digital sovereignty and control over one's own data depend on domestic capabilities and appropriate resources. Only a few middle-income countries would be in a position to successfully implement comprehensive homegrown AI strategies without resorting to international cooperation.

Global governance initiatives may seek to address these challenges by establishing innovative mechanisms to make research facilities and key infrastructure assets available, such as cloud services or supercomputers under multilateral jurisdiction. A collaborative public-private partnership could grant access to AI resources to experts from developing countries.

Future international governance institutions should include programs specifically designed to promote the peaceful uses of AI, local ownership of research and development, technology diffusion, and on-the-ground implementation in poor countries.

Joint research, open access, capacity-building, and a fair and equitable distribution of benefits are essential to leave no one behind.

Unless we take the concerns of the majority of the world’s population seriously, reaping the rewards of the AI revolution will be a privilege confined to a minority, or worse still, controlled by a few hands.

Eugenio Vargas Garcia, Tech Diplomat, Deputy Consul General of Brazil in San Francisco, Head of science, technology and innovation, and focal point for Silicon Valley. Former senior adviser to the President of the United Nations General Assembly in New York, 2018-2020.

Promoting Participation of Developing Countries in AI Governance and Sustainable Development

LU Chuanying (鲁传颖)

The development of artificial intelligence (AI) technology is creating opportunities and challenges for economic and social development for countries around the world. For developed countries, the opportunity lies in utilizing AI technology to drive industrial transformation and upgrading, empowering high-speed economic growth. For developing countries, the main opportunities lie in using AI technology to address developmental challenges. For example, in the field of healthcare, on the one hand, AI can assist doctors with data analysis, facilitating a better understanding and analysis of patients' conditions, thereby improving the healthcare capabilities of developing countries. On the other hand, from a broader perspective, AI can push developing countries to establish citizen consultations and healthcare assurance systems, enhancing the medical standards of developing nations. Additionally, AI provides numerous development opportunities to developing countries in areas such as education, infrastructure, agriculture, energy, and the environment.

However, the current dividends of AI development have not yet reached developing countries. According to the Government AI Readiness Index 2022 published by Oxford Insight, developing countries face significant gaps in terms of technological infrastructure, innovation capacity building, and governance capabilities within government, leading to a noticeable North-South divide in AI development.

Given this situation, when considering AI governance, efforts should be made to promote the participation of developing countries and advance the United Nations (UN) Sustainable Development Agenda in two aspects:

First, it is necessary to further leverage the role of international organizations such as the United Nations in implementing the sustainable development agenda. Currently, international organizations such as the UN are gradually realizing that AI governance needs to benefit more developing countries. For instance, the UN Conference on Trade and Development released a report titled "Technology and innovation for cleaner and more productive and competitive production" discussing facilitating technology transfer to developing countries through official development assistance, trade, and foreign direct investment. It also explains the operational mechanisms within the UN system and explores methods of using technology and innovation to realize inclusive and sustainable development. The International Monetary Fund conducts special studies on the impact of AI on developing economies, and the OECD AI Policy Observatory tracks the deployment of AI in developing countries, proposing recommendations for digital transformation with working papers. These international measures show a positive trend but require further implementation.

Second, China should actively participate in AI governance and speak up for developing countries. In reality, developing countries are in a weak position on the establishment of international rules for AI, partly due to the lagging development of AI in many developing countries as well as largely due to a lack of motivation and awareness to participate in international rule-making. In response to the continuous strengthening of the international rules system for AI by Western countries, developing countries should quickly awaken, increase their participation, and contribute their wisdom.

As a representative of non-Western countries and developing countries, China is one of the few developing countries actively voicing its stance on international AI rules. Therefore, China should take on the responsibility of energetically promoting deep cooperation on AI within the Global South and urging their active participation in the development of the AI industry.

China should prioritize options that promote the research, development, and application of AI technology by Global South countries and actively engage in technological and industrial cooperation with important countries and regions. Special attention should be paid to potential markets in Southeast Asia, the Middle East, and other regions. China should actively try different cooperation plans, accumulating beneficial experiences in AI development and governance from the perspective of Global South countries. China should actively promote the participation of Global South countries in the global AI industrial chain and value chain, enhancing through practical efforts their overall influence and discourse power in construction of international rules. In addition, China has abundant practical experience and precedent in utilizing AI to promote economic and social development, such as with smart cities, smart healthcare, smart education, and smart agriculture. China should actively share its knowledge, experience, and resources in the field of AI with Global South countries, thereby contributing to economic and social development and construction of the governance system.

Chapter 3: AI Governance for Developing Countries and Advancing Global Sustainable Development

Lu Chuanying (鲁传颖), researcher at the Shanghai Institutes for International Studies, Secretary General of the Cyberspace International Governance Research Center and Deputy Director of the Institute for Public Policy and Innovation Studies, specialized in cyberspace governance and cyber security.

AI Supply Chains and Geopolitics : Co-governance with the Global South

Marie-Therese Png

It is increasingly clear that countries and communities in the economic power centres of the Global North are best positioned to profit from benefits afforded by AI R&D. Meanwhile, costs are carried by those already disproportionately disadvantaged by socio-economic, geopolitical and trade dynamics. Inclusive AI governance initiatives aim to address such distributional inequalities, but have yet to incorporate into their analysis the underpinning structures. For example, AI industries are embedded in incentives structured towards financial, military, natural resource, and geopolitical advantage that marginalise the material safety of a Global Majority.

An important minority of leaders in AI governance initiatives recognise their responsibility to ensure Generative AI deployment and regulation do not lock in intra-national and international inequalities. In addition, they understand the potential strategic advantages of globally inclusive efforts - for consensus building, governance effectiveness, and geopolitical stability. They understand that extreme power imbalances between different regions can have long-term detrimental and destabilising effects at a global level - competition and conflict, and ways in which political uncertainty, trade wars, sanctions and international unrest disrupt supply chains and innovation.

A globally representative AI governance process also provides a picture of world politics that is empirically grounded. To quote Albert et al. (2020), global governance omits "the imperial and (post)colonial past and present of international relations, thereby presenting a theoretical picture of world politics that is deeply embedded in Eurocentrism and therefore exhibits serious theoretical and empirical flaws." Action-oriented discussions with Global South stakeholders - especially civil society and laypeople - based on their needs, demands and goals enables a scoping of leverages, barriers, and gaps to create informed strategies.

This, and other forms of co-governance (Brito et al., 2021; Png, 2022), afford an understanding of alignment discrepancies between intended and real outcomes of current governance processes, in order to iterate and materially serve large segments of the global Population.

Gaps in international AI governance that Global South stakeholders can provide valuable empirical evidence towards, include the negative impacts of the AI industry's reliance on normalised, but exploitative practices. The AI industry relies in part on large-scale extraction and monetisation of data, cheap digital labour, and the extraction of minerals, metals, water, and land to build hardware and physical information infrastructure. The procurement of these commodities must be examined for exploitative practices, especially given that many jurisdictions in the Global South still lack pre-existing safeguards and regulations around data protection, ownership and monetisation, digital platforms, compute, data centres, data flows, labour practices, and natural resources in complex multi-country supply chains (Veale et al., 2023). This leaves countries, and their populations, systemically more vulnerable to risks emerging from AI development. This is further compounded by infrastructurally developed states depending on and incentivised to extract from resource rich countries who are weaker members of the world market economy in ways that keep them underdeveloped (Rodney, 1972). This is likely to become more acute as Generative AI systems advance in capabilities, increasing the demand for these commodities. This requires the development of safeguards and regulations that are sensitive to these dynamics, and understand their contribution to aforementioned political uncertainty, international unrest, and supply chain volatility, that ultimately disrupt long term innovation and its benefits.

Marie-Therese Png, an AI Ethics expert and PhD candidate at Oxford University's Oxford Internet Institute (OII), earning a DPhil in Information, Communication & the Social Sciences. Founder of Implikit and previously Technology Advisor leading on the design and implementation of the UN Secretary-General's Digital Cooperation Office.

What AI Oversight Can Learn From Carbon Emissions

Charlotte Siegmann and Daniel Privitera

A global AI ecosystem that is safe for humans is a global public good: individuals cannot be excluded from benefiting from it, and one individual's benefiting from it does not reduce its availability to others. This makes it likely that, without countermeasures, AI safety will be underprovided by the market and by nation states. In an era of rapidly advancing AI capabilities, such market failures expose the global community to potentially severe risks affecting the whole world population.

However, market failures in global public good provision like clean air or a stable climate are a widely-studied phenomenon. We can learn from the failures and successes of addressing these partly analogous contribution problems, such as carbon emission reduction, to avoid market failures in the case of AI. Market-based international treaties might be particularly adept for AI governance, allowing to **broker various win-win deals, reward frontrunners, and punish violations**.

There are several potential AI market failures that might need to be addressed globally. In the absence of any agreements:

1. **Nation states might spend less on AI safety and fairness vis-à-vis AI acceleration than what would be optimal** for everyone on earth. Similarly, nation states might allow the hasty deployment of immature and dangerous AI technology to outpace competing firms or nation states.

2. **Nation states might fail to ensure that enough resources flow to highly beneficial AI technology** that the market fails to value appropriately, such as AI tools for the health sector and medical research, for mitigating climate change, or for empowering underrepresented groups.

3. **Nation states might fail to appropriately coordinate with other countries** about the goals and values to which domestic, powerful AI models get aligned. Such a norm of non-coordination would likely leave every nation state worse off than they would have been under a coordination norm.

While nation states should be responsible for the implementation of AI governance, international treaties are necessary to curb free-riding.

Each of the problems described above represents a classical market failure: Some global externalities (both positive and negative) of an action are not adequately reflected in the cost that nation states undertaking this action incur. A standard solution to such market failures consists in agreements that price in these externalities, creating incentives that are better aligned with everybody's joint interest. In the case of AI, this could take various forms:

1. **Taxation-Empowered Differential Development:** An international protocol could commit nations to monitor and track AI deployment compute. Each country would commit to subsidizing and taxing various use cases of AI differently (depending on their social cost or benefits), e.g., via compute or data usage or company revenue taxation. Moreover, countries could commit to at least using a specific portion of domestic AI research funds and domestic AI compute (whether government or company-owned) for socially beneficial purposes (which are flexibly defined by the countries) or capping the size of the AI training runs. *Analogies: Kyoto Protocol and European Emission Trading System.*

2. **Verification and Penalties:** The commitments by nation states under a Differential Development framework would need to be enforced by the international community. Various commitments such as appropriate cybersecurity measures, spending on AI safety and fairness research, and investment in "AI for good" could be verified. Adherence could be rewarded and violation penalized. In the future, countries could also jointly set up a central AI compute cluster that all participating countries would have access to. The degree (quantity) and conditions (price) of access for each country could be tied to its adherence to jointly defined rules. *Analogies: WTO sanctions, Basel III Standards, climate change proposals, arms control agreements.*

3. **Benefit-Sharing Commitments:** Moreover, countries could agree to share the benefits of domestically developed AI globally. This would reduce redundancies, incentivize nation states to not race ahead, and allow more countries to benefit from the technology's potential huge advantages. *Analogy: Nagoya Protocol on Access to Genetic Resources*

Given the rapid pace of AI development, the world needs to coordinate quickly to avoid large-scale risks. While several challenges need to be addressed, market-based mechanisms like the ones above offer three advantages: First, they provide actual financial and economic incentives for nations to abide by commitments. Second, for the most part, national regulatory capacities can be used, and the mechanisms can be adapted over time since, most of the time, governments rather than the international bodies would do the regulation. And third, we can learn from the successes and failures of existing global public good institutions and market-based mechanisms. For these reasons, beginning to experiment with, and implementing, market-based agreements could help the world reap the benefits of fair and safe AI development while safeguarding against the risks.

Charlotte Siegmann, founding member of the Center for AI Risks & Impacts and PhD student at the Massachusetts Institute of Technology. Focused on the potential global diffusion of AI policy, AI regulation, technical AI safety, and the economics of AI governance. Previously an economist at Oxford University.

Daniel Privitera, founder and Executive Director of the KIRA Center for AI Risks & Impacts and a DPhil candidate at the University of Oxford. Holds an MPhil in economics from the University of Oxford.

How can AI Governance Promote Global Economic Growth and Sustainable Development?

LIAO Lu (廖璐)

The governance of artificial intelligence (AI) holds significant potential importance for nations worldwide, particularly with regard to pursuit of high-quality economic and social development and the realization of the United Nations (UN) Sustainable Development Goals (SDGs).

First, AI governance is poised to robustly stimulate economic growth. The widespread application of AI technology can significantly enhance productivity, aiding businesses in reducing costs and increasing competitiveness. This is especially crucial for developing countries and regions where economic growth is often relied upon to alleviate poverty and improve standards of living for the people. Through applications in intelligent manufacturing, automated agriculture, and digitized services, AI can free up labor, boost productivity, and create greater employment opportunities for these countries, attracting more investments, and fostering economic diversification.

Second, AI governance can improve social services and infrastructure. Providing high-quality education, healthcare, and infrastructure to the population is often a challenge in developing countries. AI, as an auxiliary tool, can be used to improve education, healthcare systems, and urban planning, offering better services to the people, enhancing efficiency of resource utilization, and reducing waste. This contributes to improved social welfare, reduced inequality, and increased support and belief in sustainable development.

Additionally, AI governance can facilitate the implementation of SDGs. The United Nations SDGs encompass various aspects such as poverty eradication, environmental protection, peace promotion, and improvement of education. AI can be utilized to monitor and assess the progress of these goals, providing data support for decision-making. It can also assist countries in better managing resources, reducing environmental impact, and thereby better achieving sustainable development.

However, AI governance comes with many challenges, including potential privacy issues, ethical concerns, and digital divides. Therefore, nations need to establish and refine robust legal frameworks and policies to ensure that the development and application of AI will not exacerbate social inequality. Moreover, nations should broaden and strengthen international cooperation and exchange, increase sharing of knowledge and experiences, share best practices, and meet the challenges together. Governments and relevant international organizations should take the lead in drafting relevant laws, regulations, and principles, establishing review and feedback mechanisms, and ensuring that AI sufficiently understands and respects differences in culture and values among diverse populations in both the development and use phases. This approach seeks to ensure that AI effectively avoids bias or discrimination and better serves the rights and interests of global users.

In conclusion, AI governance holds immense potential for nations worldwide, particularly for the high quality economic and social development of developing countries and for implementing the UN SDGs. By implementing rational policies and governance measures, AI can become a powerful tool in achieving these goals.

Liao Lu (廖璐), Senior Programme & International Cooperation Manager at the Beijing Academy of Artificial Intelligence (BAAI) focused on AI international cooperation and young scientists community building.

AI Governance for the Global Majority – Southeast Asia, a case in point

Lyantoniette Chua

Southeast Asia (SEA) is widely recognized as one of the largest data pools globally, the very resource that fuels AI training models which is the base for developing new global and essential AI platforms. AI's potential economic contribution could be between $10-$15 trillion by 2030. At large, the SEA region is also the global epicentrum of growth in this decade. In ASEAN, various countries have made initiatives and strategic policies to orchestrate AI development according to each country's goals and targets.

However, the vast promises of AI also come with inherent risks and potential ramifications. The misuse or malicious use of AI can lead to catastrophic consequences, especially given the diverse sociopolitical contexts and digital disparities across SEA nations.

Introducing the Capability Justice - Context Based approach as the basis for the inter-regional governance of AI in Southeast Asia and neighboring regions, through a document submitted to the United Nations Office of the Secretary-General's Envoy on Technology (UN Tech Envoy) titled: Stamping Southeast Asian Voices on Global Governance of AI: A Seven-Point Plan for the UN Advisory Board on AI. This document is the abridged version of the first edition of The Ambit's Whitepaper series on Southeast Asia and AI Governance.

The white paper will be released according to the editions below:

- 2023: First Edition - Indonesia, Philippines, and Singapore (Cluster 1 of 3)
- 2024: Second Edition - Malaysia, Thailand, Vietnam, Myanmar (Cluster 2 of 3)
- 2025: Third Edition - Cambodia, Brunei, Timor Leste, Laos (Cluster 3 of 3)

In The Ambit's Global Network submission to the UN Tech Envoy, the 7-point plan carries the following key actions:

1. Establish inclusive and diverse High-Level Experts at the Inter-regional Level to monitor the impact of AI on society.

2. Endorse and support the upcoming 2024 Southeast Asia AI Governance Roadmap Hackathon.

3. Dedicate an institute on AI Governance for the global majority (the Global South) or non-G20 nations and conduct an AI Development-Deployment-Governance Readiness Mapping.

4. Carry forward 2023-2028 Strategy for the Inter-Regional Body on Responsible AI Governance with an executive order from the national governments.

5. Carry forward and establish platforms for Cross-Border Collaboration to encourage exchange and knowledge sharing on AI governance, research, and development among neighboring countries in Southeast Asia.

6. Carry forward the AI Development Governance Capacity Building Initiative.

7. Support the Capability Justice - Context Based approach as the basis for the inter-regional governance of AI in Southeast Asia and neighboring countries.

Southeast Asia must not navigate this journey in isolation. International collaboration and knowledge sharing will be critical to collectively address the global challenges posed by AI, ensuring that its development is inclusive and considers the needs and aspirations of all nations. AI is not bound by borders, and its governance should reflect this reality. Beneath the surface, Southeast Asia stands at a crossroads where responsible AI deployment can foster economic growth, social progress, and equitable development. By embracing the promises of cutting-edge AI ecosystem development while proactively mitigating its global challenges, the region can pave the way for a future where technology serves as a force for good, leaving no one behind globally. It is a journey that requires the collective efforts of governments, industries, academia, and civil society, and one that holds the promise of a brighter, more equitable, and prosperous future for all.

Lyantoniette Chua, Policy Group Coordinator and Fellow at the Center for AI and Digital Policy, Washington DC. Recently founded The Ambit and sits as its Global Convenor, as well as the council Co-Chair of its national founding chapter The Ambit Philippines. Previously a Vice-Chair of the IEEE Tech and Concentration of Power Committee under The IEEE Global Autonomous/Intelligent Systems Ethics Initiative.

Chapter 4: AI Governance from an Engineering Perspective

Understanding Model Capabilities is a Global AI Governance Priority

Nathaniel Sharadin

As capable, general-purpose machine learning (ML) models improve and proliferate, there is increased public interest in evaluating models in order to effectively govern their deployment and development. For instance, the UK government's recent whitepaper calls for a "toolbox" of techniques that can "measure, evaluate, and communicate" model capabilities, noting that "the extent of [large models'] capabilities" is an open research question. The Biden administration has called for "independent" evaluation by researchers with "unfiltered access" to models with unknown capabilities. Separately, large model developers, including Google, OpenAI, and Meta have publicly, voluntarily agreed to advance research into "capability evaluations" and to "developing a multi-faceted, specialized, and detailed" regime for evaluating (and reporting) the capabilities of models.

It isn't all rhetoric and voluntary agreements: recent U.S. law provides for the establishment of "testbeds, including virtual environments" to examine machine learning systems. The EU AI Act, now adopted by the European Parliament in draft form, goes further. Compliance with the AI Act requires that model developers provide a "[d]escription of the capabilities" of (especially general-purpose) models. And in Asia, China has released preliminary rules governing generative AI, the implementation of which will likely require developers (or officials) to evaluate model capabilities; ASEAN is preparing guidance for their member nations that will recommend model evaluations. And there's growing international multilateral consensus on governance frameworks for foundation models that target "capability thresholds."

So, there is widespread agreement that in order to effectively govern the development and deployment of ML models we need robust, systematic evaluations of models' capabilities. It is therefore striking that there is no systematic conceptual framework for deciding what ML models are in fact capable of doing. This is despite the fact that claims about ML models' capabilities are ubiquitous. Large model developers (and their critics) make specific claims about what particular models are (or are not) capable of doing — passing the bar exam, effectively deceiving humans, generating misinformation, producing hate speech, etc. — and we are told that model capabilities may be dangerous, harmful, beneficial, emergent, autonomous, or novel. Models are said to have capabilities in chemistry, medicine, programming, hacking, war-making, and a huge range of other domains. But despite this, there is, again, no systematic account of what it is for an ML model to have a capability, or how, precisely, claims about model capabilities are to be decided. Talk of capabilities is simply not interrogated.

This is a serious gap in our joint understanding of this new technology, and this gap is an important barrier to effective governance. For instance, large actors cannot cooperate to limit the scope of model capabilities unless they antecedently agree on what counts as evidence of a model's having a capability in the first place. As a precursor to the important work of establishing both domestic and international oversight of the development and deployment of capable machine learning systems, we therefore require a systematic framework for deciding claims about what models are capable of doing. Developing such a framework should be a governance priority.

Nathaniel Sharadin, Assistant Professor (Philosophy) at the University of Hong Kong and a Research Affiliate at the Center for AI Safety, a San Francisco-based non-profit aimed at reducing societal-scale risks from AI. Currently working on two AI-related projects: one on understanding (and evaluating) large, frontier ML model capabilities and another on the nature, value, and importance of artificial achievements.

AI Safety and Security Governance, Global Cooperation, and Agile Iteration from a Standardization Lens

MA Chenghao (马骋昊), GAO Wanqi (高万琪), and FAN Siyu (范思雨)

In today's era of continuous emergence of intelligence, technological potential and hidden risks seem to coexist. On the one hand, artificial intelligence (AI) continues to make progress and breakthroughs in downstream tasks, updating and elevating the public's awareness and expectations. On the other hand, AI's ethical, safety, and security issues such as privacy leaks, bias and discrimination, responsibility attribution, and technological misuse are also continuously emerging.

In this situation, standardization will be a practical and feasible dimension of thinking and governance. Internationally, standardization work on AI safety and security governance has been carried out in various dimensions by standard organizations such as the International Organization for Standardization (ISO), International Electrotechnical Commission (IEC), International Telecommunication Union (ITU), and Institute of Electrical and Electronics Engineers (IEEE). A certain level of international consensus has already been formed on topics including vocabulary, foundational systems, and risk management frameworks. Representative examples include the Trustworthiness working group (WG 3) established by the ISO/IEC JTC 1/SC 42 AI Subcommittee and IEC/SEG 10 Ethics in Autonomous and Artificial Intelligence Applications.

In 2018, China established the National Artificial Intelligence Standardization Overall Group, which includes the Artificial Intelligence and Social Ethics and Morality Standardization Research Group. Subsequently, the China National Information Technology Standardization Committee's AI Subcommittee (SAC/TC 28/SC 42) established the Trustworthiness Research Group, and the National Information Security Standardization Technical Committee (SAC/TC 260) concurrently initiated the research and development of domestic standards relating to information security.

Currently, China has already developed a set of AI ethical norms and related standards, comprehensively laying out technical research and applications related to AI ethics governance. There are already key standards in the drafting stage on issues such as AI management systems, risk management capability assessments, and trustworthiness. These key standards have clear boundaries and scope in the field of AI safety, security, and ethics governance, and they can apply broad and abstract ethical principles to practical technological research, guiding the compliant development of the industry.

However, different countries, regions, and industries have different requirements for AI safety, security, and ethical issues. Most current standardization work mainly appears in the form of guidelines, frameworks, principles, and criteria, while more specific, implementable technical standards are still in the exploration and early research stages, awaiting more researchers to join and collaborate.

Therefore, using standardization work as a starting point, we propose several initiatives, in hope of offering a modest suggest that inspires greater contributions, and further exploration:

- The first proposal is to iterate upon agile updating mechanisms at the criteria level for AI technology. Construct a global researcher community based on data, cases, and experiments, continuously updating and refining the content and scope of AI safety and security governance guidelines in real-time.

- Second, we call for the research and development of special standards in sensitive application areas such as education and healthcare to promote the implementation and application of high-level guidelines. Conduct extensive social experiments and interdisciplinary discussions, build specialized governance datasets and governance technologies for AI applications in various industries.

- Third, we suggest the establishment of an International AI Alignment and Governance Innovation Demonstration Zone in Shenzhen. We call on global AI companies and research institutions to participate together, and within a certain scope validate how scientific and how operationalizeable are relevant alignment methods, standards and norms, governance tools, data sharing mechanisms, and other such content.

In summary, we collectively envision a more beautiful, safe, and aligned artificial general intelligence future.

Ma Chenghao (马骋昊), China Electronics Standardization Institute (CESI) Engineer, Chair of IEEE/C/AISC/LSDLM, editor of five national AI standards, chief editor of many AI national and industry standards.

Gao Wanqi (高万琪), CESI South China Branch AI industry researcher, primarily responsible for AI safety and security governance and industrial policies. Previously participated in editing and researching many AI industry policies.

Fan Siyu (范思雨), CESI South China Branch industry researcher, teacher for AI popular science classes at the Digital Information Products Standardization National Engineering Research Center.

AI Governance — Another Tower of Babel

WANG Jun (王俊), NA Diya (娜迪娅)

In 2022, the development of generative artificial intelligence (AI) injected vitality into the field of AI. The emergence of ChatGPT is seen as the starting point of artificial general intelligence (AGI) and a turning point towards strong AI, sparking a new round of AI revolution. AI development seems to have found its mainstream narrative.

However, technological innovation has brought governance challenges, and what we face is not immediately within grasp, rather it is a future beyond the reach of imagination. For disruptive AI technologies, governance approaches need to be proposed from a global perspective. Countries, regions, and relevant companies, experts, scholars, societies, and publics should dismantle separation, address the divisions within the rules, and collectively discuss the challenges we must face.

Based on Nancai Compliance Technology Research Institute's continued observation and reports on AI, we have summarized several prominent issues facing global AI development:

1. In the global competition over AI governance discourse, issues of interoperability of rules are gradually becoming apparent.

The application of AI, exemplified by ChatGPT, has triggered a surge in industrial and technological development. While both domestic and foreign technology giants are increasing their investments in technology and capital, it has also raised regulatory concerns.

It is apparent that various countries and regions have begun taking measures to strengthen regulation. The European Union's "AI Act" is in the final negotiation stage, and the U.S. has announced a series of new measures related to the use and development of AI in the country.

China's "Interim Measures for the Management of Generative AI Services," released in July, became the first policy document in the world specifically targeting generative AI.

On the one hand, there is an AI development race, and on the other hand, there is a competition over the formulation of regulatory discourse. Different regions have different drivers, and each country or region are formulating regulatory policies based on their own AI development situation, leading to the problem of rule interconnectedness. There is already a practical issue wherein the lack of unified regulatory standards has resulted in problems with applications of AI models.

With the implementation of legislation, regulation, and policies in more countries and regions, if rules appear fragmented and divided, it will be detrimental to the long-term development of AI and hinder achieving global consensus. Globally, there should be strengthened coordination on rules, after all the new challenge we face is the human-machine challenge.

2. Distant concerns—the issue of value alignment

Issues of safety and alignment have arisen accompanying the development of AI. The importance of this issue can be seen in Sam Altman's vigorous and proactive call for regulation and emphasis on value alignment at OpenAI.

Especially when a technology is extensively applied in various fields, previous, scattered issues like discrimination and lack of safety will likely become more concentrated.

However, value alignment is no easy task, even if there is agreement on principles such as "helpful, honest, and harmless," the specifics can be distorted over space and time. In addition, firm value judgments are not suitable for AI, as no one possesses the authority to pass judgment about values.

Stepping deeper into cyberspace, all realities can be mirrored within AI systems. The power structure behind code management patterns are both concealed and profound. AI alignment resembles a race against time, requiring the discovery of solutions that ensure the controllability of AI.

3. Will AI development bring about a digital dividend or exacerbate divides?

What will AI development bring to the society and public? Higher work efficiency, a more convenient life, or exploitation that is better concealed?

We discuss this issue from three dimensions.

1. Are enterprises and capital becoming more concentrated? Despite the current "battle of a thousand models," the primary controllers are still companies such as

Microsoft, Google, Two Sigma, OpenAI, etc. The shadow of the "digital capital empire" in the platform economy era still exists and is more prominent in the AI era.

2. The future widening gap between different countries. Since AI systems are often built by companies in developed countries, developed countries therefore mainly control the core of the industrial chain, but data processing and other industries in this industrial chain are mainly conducted by developing countries. Many studies abroad point out that this will further exploit the human capital and resources of developing countries, exacerbating wealth inequality.

3. It becomes more difficult for individuals to confront the system. Code is law, if the rules deviate from a human-centric perspective, how can individuals provide feedback or resistance? Additionally, how can dividends rather than fears be shared during the application of AI systems?

For these issues, we propose the following measures:

1. Strengthen dialogue and communication mechanisms

As AI systems become embedded in human society, countries and regions should strengthen their interactions with each other, observe the effectiveness and impact of different regulatory measures, seek common ground while respecting differences, and strive for consensus.

Non-governmental organizations, civil society organizations, think tanks, and others should play their unique roles and actively promote global exchange and communication.

2. More inclusive, diverse, and transparent safety rules for AGI

The problem of value alignment is essentially an ancient human problem with no correct answer. "Culture is relative, morality is absolute... In a multi-civilization world, the constructive path is to reject universalism, accept diversity, and seek commonality."

Different industries, fields, and organizations should participate in discussions, carefully considering current and upcoming situations. This includes how to annotate data during the training dataset process, training to provide timely feedback on hidden biases, and balancing the value preferences of different countries and users... to establish a system that is more fair and inclusive.

In May, the first large language model governance open-source Chinese dataset, 100PoisonMpts, was released domestically. Over ten renowned experts and scholars became the initial engineers for annotating the first batch of "100 bottles of poison for AI." Each annotator proposed 100 tricky questions inducing biases and discriminatory responses, annotating the large model's answers, completing an offensive-defensive "poisoning" and "detoxification" process. This is a highly beneficial attempt, and similar experiences should be shared and exchanged more frequently.

3. Promote open-source availability of high-quality datasets

High-quality training datasets can alleviate AI discrimination problems to some extent and promote diversity.

Currently, there is a problem of depletion of high-quality training datasets globally, and the reality is that there is a shortage of Chinese language corpus data.

The Guangdong–Hong Kong–Macao Greater Bay Area (GB) has advantages in the massive scale of its data and richness of application scenarios, and its data factors market is expanding continuously. It should leverage its advantages, fully tap into the value of data, and upon the basis of data compliance, further promote open access to public data, advance the construction of multi-modal public datasets, and create high-quality Chinese language corpus data.

4. Build an industrial chain for data upstream, midstream, and downstream

The AI industry chain is dispersed and complex. Seizing the opportunity of data element market construction and vast prospects in the industries of data services, data processing, etc., the Guangdong–Hong Kong–Macao Greater Bay Area, while strengthening international cooperation, should build an autonomous and controllable data industrial chain. With the rapid development of model training materials, such as AI training data and automated decision-making, the GBA should take the construction of the data element market as an opportunity, effectively carry out various data services, enable information hosting by individuals, revitalize the compliant and safe circulation of personal information, and enable individuals to share the dividends of AI. At the same time, focusing on the upstream and downstream of the industrial chain for model training, such as data cleaning and data trading, will accelerate develop, promoting the overall development of the data element market.

Wang Jun (王俊), first researcher at the Nancai Compliance Technology Research Institute. Focused on the convergence between technology and law, researching frontier issues including anti-monopoly and anti-unfair competition, personal information protection, and interconnectivity.

Na Diya (娜迪娅), Deputy Dean at the Nancai Compliance Technology Research Institute, Deputy Dean of the Maritime Silk Road Research Institute, focused on personal information protection, data security, data elements markets, etc.

Singapore's approach to governing Generative AI

Denise Wong

Generative AI (GAI) presents huge opportunities for different domains and applications, from enhancing user experience to productivity gain. It also comes with risks that people are concerned about:

1. **Mistakes and hallucinations.** GAI models can make mistakes, and its "hallucinations" can be deceptively convincing or authentic.

2. **Privacy and confidentiality.** As GAI models have a tendency to memorise training data, adversaries may be able to reconstruct sensitive data by querying the model.

3. **Disinformation, toxicity and cyberthreats.** Dissemination of false content such as fake news is becoming increasingly hard to identify due to convincing but misleading text, images and videos, generated at scale. Toxic content may also be propagated by GAI.

4. **Copyright challenges.** GAI models could be trained on data that includes copyrighted material, resulting in the creation of unauthorised derivative works.

5. **Embedded bias.** AI models can amplify the biases in the training dataset, potentially leading to biased outputs in downstream applications.

6. **Values and alignment.** AI systems can be misaligned with human values and goals, leading to potentially dangerous outcomes.

To build trust in GAI and ensure that GAI is developed and used responsibly, all stakeholders, including governments, industry, academia, and civil society have a role to play.

Currently, there is a lack of international consensus and alignment on principles, frameworks, standards and tools to govern GAI and help industry deploy GAI responsibly.

To contribute to international discourse on responsible GAI, Singapore has published a **Discussion Paper** – Generative AI: Implications for Trust and Governance that recommends a practical, risk-based and multi-stakeholder approach towards the governance of GAI:

1. **Model Development and Deployment:** Model developers should be transparent about how their models are developed and tested. Policymakers can support through facilitating the development of standardised evaluation metrics and a corpus of tools and capabilities.

2. **Assurance and Evaluation:** Third-party evaluation and assurance is crucial to enhance credibility and trust. "Crowding in" open-source expertise will be critical in growing a vibrant ecosystem for third-party testing of AI systems. As countries seeks to ensure that AI models are aligned to their unique values and AI governance principles, and companies train their AI models on specific datasets, industry collaboration and customised test benchmarks will be important.

3. **Safety and Alignment Research:** Policymakers need to invest to accelerate safety and alignment research, to enable interpretability, controllability and robustness. This effort should also nurture centres of knowledge in Asia and other parts of the world, to complement the ongoing efforts in the US and EU.

Singapore also aims to contribute tools to help companies test and evaluate GAI. Singapore has set up **AI Verify Foundation** to open-source AI Verify Minimum Viable Product, an AI governance testing framework and software toolkit, currently for testing discriminative AI. AI Verify Foundation seeks to tap on the global open-source community to expand AI Verify to have the capabilities to evaluate GAI applications.

Denise Wong, Assistant Chief Executive, Data Innovation & Protection Group - Singapore Infocomm Media Development Authority (IMDA). Scope of work includes developing forward-thinking governance on AI and data, driving a pipeline of AI talent, promoting industry adoption of AI and data analytics, as well as building specific AI and data science capabilities in Singapore.

AI Alignment by AI Democratization

Elizabeth Seger

Democratizing AI Development via Open-Source

Though AI democratization is a multifaceted concept, the term is very often used to refer to the democratization of AI development. Democratizing AI development is about helping a wide range of people contribute to AI development processes (Seger, Ovadya, et al., 2023). When people with diverse life-experiences and geographic, economic, and cultural backgrounds are enabled and supported to participate in AI development processes, the AI products developed are more likely to be well aligned with diverse user needs, more so than if development were concentrated within a few leading labs in Silicon Valley.

For the past 30 years open-source communities have played an important role democratizing software development, and now AI development, to serve diverse human interests and needs. Open-source AI development refers to the collaborative process of creating AI models and tools that are made freely available for anyone to view, use, study, modify, and distribute, fostering transparency and community-driven innovation. For example, the large language model BLOOM was developed over the course of a year by a global coalition of over 1000 volunteer AI developers yielding an LLM, functional in 46 languages (BLOOM, 2022).

In addition to developing and sharing open-source AI models, open-source communities engage in further, proactive efforts to democratize AI development, for instance, through education and outreach and even pooling compute resources to share costs of model training.

The benefits of open-source development to democratizing AI development are significant. However, open-sourcing also carries risks which must be considered in balance (Seger, Dreksler, et al., 2023). In particular, the consequences of malicious use are becoming more significant as model capabilities increase (Anderljung & Hazell, 2023; Shevlane et al., 2023). With access to model code and weights, malicious actors can disable safeguards against misuse and possibly introduce new dangerous capabilities via fine-tuning (Qi et al., 2023; Rando et al., 2022). The decision to open-source is also irreversible; there are no take-backs if harms emerge. Decisions to open-source should therefore be made with care.

Democratizing AI Governance

Where the risks of democratizing AI development are high, the democratization of AI governance serves as a second mechanism for helping to align AI development and deployment decisions to wider public interests and values.

Democratizing AI governance is about distributing influence over decisions about AI to a wider community of stakeholders and impacted populations (Seger, Ovadya, et al., 2023). AI governance decisions involve balancing AI related risks and benefits to determine how and by whom AI is developed, distributed, used, and regulated.

There are various options to explore for democratizing decision-making about AI:

Public participation and deliberation. One set of possibilities involves directly eliciting public input via participatory or deliberative democratic processes. These processes might leverage online tools (possibly AI-enabled) like Pol.is to solicit and synthesize public input into complex normative decisions (Ovadya, 2023). OpenAI recently launched a "democratic inputs to AI" grant program to experiment with setting up democratic processes for deciding what rules AI systems should follow (Zaremba et al., 2023). The Collective Intelligence Project (CIP) is similarly experimenting with 'alignment assemblies' to help identify collective values for reflection in AI behavior, in addition to navigating other complex value-laden issues such model release decisions and describing acceptable risk thresholds (Collective Intelligence Project, 2023).

Institutional structure. Large labs could also introduce organizational structures that are more democratic in nature, for example, by implementing oversight boards that are democratically elected or selected by sortition. They might also incorporate as public benefits companies to provide a clearer legal standing for making decisions to maximize public benefit even if doing so conflicts with shareholder interest.

Democratically-informed regulation. Another option is to support regulation that is developed in response to deliberative processes involving developers, open source communities, academia, and civil society to reflect diverse stakeholder interests.

Conclusion

It is very difficult to pinpoint precise values to which AI behavior and development decisions should respond. Instead we might look to employ fair processes for collecting and integrating diverse stakeholder inputs. Towards this end, one option is to involve many more people directly in AI development processes through open-source development. Open-sourcing does, however, carry risks. Where the risks are high, AI alignment might instead be pursued involving diverse stakeholders in democratic decision-making to inform consequential decisions.

Elizabeth Seger, Research Scholar at the Centre for the Governance of AI and Research Affiliate at the AI: Futures and Responsibility Project at the University of Cambridge Centre for the Study of Existential Risk.

Engineered Wisdom for Learning Machines

Brett Karlan and Colin Allen

Recent successes of deep neural networks have led some to believe that the age of artificial general intelligence (AGI) is not far off. Whether such a future is indeed imminent is anyone's guess, though we remain skeptical.

State-of-the-art deep neural networks are impressive in their ability to sort through massive amounts of data and detect patterns. They are, however, surprisingly brittle in their responses. For instance, language models which try to generate plausible text based on user input show a significant decline in the quality and intelligibility of their outputs as the requested amount of text gets longer and they are unable to monitor their own inconsistencies. Another pitfall of state-of-the-art deep neural networks is their opacity. It is often simply not possible to know what information a neural network used to make a decision, how it processed that information, or even where the information is stored in the network. This opacity is one factor contributing to users overestimating the capacities of AI. Another factor is that users do not rigorously test the limits of these systems.

How can we make better algorithm-aided decisions? What should we be aiming for as a goal in human-machine interactions? We argue that the concept of practical wisdom is particularly useful for organizing, understanding, and improving human-machine interactions. Practical wisdom refers to a suite of knowledge, understanding, and skills aimed at coming to truths about a domain and making better decisions in that domain. Our conception of practical wisdom foregrounds two important components: (1) the meta-cognitive awareness required to come up with a rational strategy for dealing with one's own limits and the limitations of the technology; (2) the breadth of understanding, knowledge, and skill required for making good judgments.

Developing practical wisdom in human-machine interactions, in particular focusing on strategy selection and context-specific meta-cognitive reasoning, represents a way of both conceptualizing the problem of brittleness and minimizing the possibility of massive error. Several recent studies have shown that wise reasoning benefits from situational awareness on the part of the reasoner: subjects tend to reason more wisely when they dissociate themselves from their own personal investments and consider their abilities and limitations from a more third-person perspective. This supports better metacognitive awareness of the limits of one's understanding of a situation.

The opacity of AI has been discussed and theorized at length. Using the framework of practical wisdom in human-machine interactions, we can see another reason why explainable AI might be valuable: because it helps facilitate the development of practical wisdom. If a decision-maker has access to more information about how a deep neural network came to the output that it did, then she will tend to make better decisions than if she is presented with the same output and no explanation.

While the individual (or group) user of AI is an important locus of analysis as the ultimate seat of decision making, it would be a mistake to assume that the practical wisdom framework cannot be extended to other aspects of human-machine interactions. One strength of our proposal is that the notion of practical wisdom can be helpful for understanding all levels of the production and use of deep neural networks and other AI technologies in important decision domains. Developing practical wisdom at all stages of the creation, design, and implementation process for AI, in turn, makes the process of exercising practical wisdom in our decisions significantly easier.

Our conceptualization of practical wisdom offers a powerful framework for understanding human-machine interactions. This framework supports better accounts of both what success in this domain looks like, and how failures can be avoided in the future. Though other approaches and frameworks might be able to make similar recommendations to the ones we make here, our framework provides a unified and coherent set of capacities that explain how and why these recommendations should be made. Further conceptual refinement will require input from psychologists, behavioral scientists, engineers, and other stakeholders.

Brett Karlan, Assistant Professor, Department of Philosophy, Purdue University. Primarily focused on the intersection of the philosophy of science (especially cognitive science and artificial intelligence) and normative philosophy (especially epistemology, ethics, and the philosophy of action).

Colin Allen, Distinguished Professor, Department of Philosophy, University of California, Santa Barbara. Specializing in Philosophy of Cognitive Science, Animal Minds, Cognitive Evolution, Machine Morality, and Computational Humanities.

Diversity, Openness, and Interaction: Principles for Training Generative AI Models

JeeLoo Liu

Generative AI models are trained on extensive datasets composed of existing texts sourced from books, articles, online sources, and other available data. The strength of LLM training lies in its ability to train AI to process information directly from natural languages and generate grammatically and stylistically excellent content. The most successful model to date, ChatGPT-4, has demonstrated lightning speed in processing information for a vast range of topics. Even though it has been criticized as making many factual errors, these mistakes should be able to be avoided in the future generation of Generative AI.

However, there are also pressing concerns with this machine learning methodology, as evidenced in the performance of models like ChatGPT:

1. The training is based on existing data, and thus the outcome does not reflect any newly developed situations or novel inputs. As Hume has pointed out: there is no guarantee that the future will resemble the past. The existing approach risks neglecting future examples, and perpetuating current biases, injustice, and wrongful discrimination present in any society.

2. The existing texts from books and articles only include those items that are electronically accessible. Thus, the outcome will always exclude ancient non-digitized texts and non-digitized texts from underdeveloped countries or marginalized cultures. We are therefore omitting invaluable perspectives represented in non-digitized texts, and dismissing inputs that are essential for a holistic human knowledgebase.

3. Data curation often happens internally and privately within companies like OpenAI, Google, and Tesla, offering no transparency and explainability to the public. Without checks and balances, products may not fairly represent general welfare or public sentiments.

4. While “human-in-the-loop” methodologies help to eliminate toxic data, inciting languages, lurid contents, biases, and other harmful contents, the immense data volume often forces outsourcing to cheaper labor markets in underdeveloped nations. This practice raised concerns of labor exploitation and subpar data curation. There is no further sanction and guidance performed by experts on human values, such as philosophers, ethicists, moral leaders, educators, as well as the general public in the given society.

To curb these problems, I suggest:

1. We need to establish a massive database encompassing ethical behaviors, philosophical insights, and moral deliberations of virtuous moral agents across history, cultures, and languages. This will help us build an ethical model to be superimposed on current LLM models. The model does not need to present a monolithic value structure or consensus in moral judgments; virtuous people would virtuously disagree with one another. The outcome has to be pluralistic to represent the diverse cultures and values globally.

2. We also need to have an open platform to poll the general public’s opinions on a variety of issues, and have the data automatically added to the curated data for refined training. This is the true form of “human in the loop machine learning” — bring all people in the loop, and not just specific groups such as the hired MTURK workers.

3. Finally, data collection and machine learning must remain open-ended, adaptable to fresh data and evolving scenarios. The learning and training must be interactive and proactive with each additional input. Humans learn from being open to others’ suggestions; machines must do so too.

Jeeloo Liu, Professor of Philosophy, Department of Philosophy, California State University Fullerton. Research areas include Philosophy of Mind, Chinese Philosophy, and Metaphysics.

Chapter 5: AI Governance from the Perspective of Companies

A Framework for Responsibly Scaling Artificial Intelligence Models

Michael Sellitto

As frontier artificial intelligence (AI) models become more capable, they will create major economic and social value, but will also present increasingly severe risks that will need to be managed. In order to manage those risks, Anthropic has designed and adopted a Responsible Scaling Policy (RSP).

Our RSP focuses on catastrophic risks – those where an AI model directly causes large-scale devastation. Such risks can come from deliberate misuse of models (for example use by terrorists to create bioweapons) or from models that cause destruction by acting autonomously in ways contrary to the intent of their designers. While AI represents a spectrum of risks that must be addressed, our RSP is designed to deal with this more extreme end of this spectrum.

Central to our plan is the concept of AI safety levels (ASL), which are modeled loosely after the US government's biosafety level (BSL) standards for handling of dangerous biological materials. We define a series of AI capability thresholds that represent increasing potential risks, such that each ASL requires more stringent safety, security, and operational measures than the previous one.

Higher ASL models are also likely to be associated with increasingly powerful beneficial applications, so our goal is not to prohibit development of these models, but rather to safely enable their use with appropriate precautions.

Thus, the RSP brings a concrete and empirical approach to identifying and managing catastrophic risks. It is designed to allow our safety research and societally-beneficial applications of the technology to continue to develop and scale, while imposing a rigorous process for measuring and mitigating risks. In cases where significant risks cannot be mitigated, we will pause further scaling of the relevant model, refrain from deploying it, or remove it from deployment until we can assure that it is sufficiently safe to continue.

By pausing development and deployment when scaling outpaces safety, we are thus incentivized to solve the necessary safety issues. If adopted as a standard across frontier labs and supported by governments, RSPs might create a "race to the top" dynamic where competitive incentives are directly channeled into solving safety problems.

Notably, the RSP is also flexible and adaptive. The RSP specifies concrete safety commitments for current (ASL-2) and near-term (ASL-3) AI systems, spanning security, training oversight, red teaming, model evaluations, and responsible deployment measures. And it commits Anthropic to iteratively define higher ASLs (starting with ASL-4) before reaching them, ensuring safety protocols evolve alongside capabilities as new, empirical information is learned over time. Thus responsible scaling policies are meant to evolve over time, maintaining their relevance far into the future.

The full policy is available at:

https://www.anthropic.com/index/anthropics-responsible-scaling-policy.

Michael Sellitto, Head of Geopolitics and Security Policy at Anthropic, an AI safety and research company. Examines the impact of artificial intelligence (AI) technologies on issues related to national competitiveness, international relations, and international security.

Shaping a Healthy and Sustainable Development Ecosystem for Large AI Models through AI Alignment

Jason SI (司晓) and Jeff CAO (曹建峰)

With the rapid development of generative artificial intelligence (AI) large models, their autonomy, general-purposeness, and ease of use have increased quickly, and large models are poised to permeate diverse sectors, offering significant economic and societal benefits. Concurrently, these large models face ethical challenges such as hallucinations, discrimination, misuse, and emergent risks. Therefore, how can large models' capabilities and activities be identical to the values, actual intentions, and ethical principles of humans, how can we ensure the safety and trustworthiness of human-AI collaboration processes? This is the crucial issue of "AI value alignment." In order to make large models safer, more reliable, and more usable, it is essential to prevent harmful outputs or misuse of the models as much as possible. This is a core task in value alignment of current large models, and both industry and research institutions have been exploring relevant technical and governance measures.

A core technical approach at present involves enhancing AI large models' understanding of human values, ethical principles, etc. through reinforcement learning. Reinforcement learning from human feedback (RLHF) has proven effective in this regard, encompassing steps including initial model training, collecting human feedback, reinforcement learning, and iterative processes. The core thinking behind RLHF is to have human trainers assess the appropriateness of model outputs, using collected human feedback to construct reward signals for reinforcement learning, thereby enhancing model performance. In practice, RLHF has demonstrated significant advantages in areas like improving model performance, increasing suitability, reducing bias, and improving safety, including reducing the likelihood of producing harmful content in the future. Nevertheless, RLHF faces challenges such as scalability issues, subjectivity influenced by trainers' preferences, and difficulties in ensuring long-term value alignment. Therefore, researchers are exploring how to transition from relatively inefficient "human supervision" to more efficient "scalable oversight," with the idea of AI supervision gaining greater emphasis. The core idea of AI supervision is to have AI models assist human trainers or autonomously evaluate the outputs of large models, which provides some helpful explorations to present AI practices.

Moreover, there are additional methods and governance measures that can be employed to ensure value alignment is achieved in large models. One way is to intervene on the training data, such as recording and identifying issues related to representativeness or lack of diversity, conducting manual or automated filtering and testing to identify and eliminate harmful biases, or creating specialized datasets for value alignment. A second approach is to construct interpretable and understandable large models. We must better understand how large models make decisions in order to realize AI alignment. Third, adversarial testing, or red team testing, involves probing models with exploratory or hazardous questions to uncover inaccurate information, harmful content, false information, discrimination, language biases, safety or security risks, and other issues, facilitating model improvements.

In conclusion, the importance of AI alignment lies not only in being an essential path for current large models but also in its significance for the future of superintelligent AI. In the race between AI and time, collaborative efforts are crucial to drive broader interdisciplinary engagement and cooperation, ensuring the healthy and sustainable development of large models, so that future, more powerful AI will continue to benefit humanity and society.

Jason Si (司晓), Dean of Tencent Research Institute, Vice President of Tencent Group, Deputy Director of the National Network Copyright Industry Research Base, President of the Shenzhen Copyright Association, visiting scholar at Stanford University.

Jeff Cao (曹建峰), senior research fellow at Tencent Research Institute, expert at Internet Society of China, a part-time researcher at East China University of Political Science and Law, and a member of the National AI Standardization Group.

Don’t Let Black Box AI Lock In Our Civilization’s Evolutionary Path

WEI Tao (韦韬)

The continuous development of large-scale pre-trained language models, such as GPT, is ushering in a new era in artificial intelligence (AI) technology. These models have already demonstrated capabilities surpassing human abilities in many natural language processing tasks, but they also face serious challenges. Currently, their most significant issues include a lack of cognitive alignment, a lack of principles, and black-box reasoning. For example, they may generate severe hallucination problems due to their inability to effectively handle "unknowns," and they may also fail to adhere to "privacy protection," leading to potential leakage of personal information. Moreover, errors in decision-making and recommendations during the reasoning process are challenging to deconstruct and verify for various reasons.

In the face of these risks and due to the unlimited energy of AI, the wrong decision making caused by these problems could be rapidly and widely executed by large models, leading to difficult to foresee and severe consequences. AI is like an immature child, but with the power to wreak havoc. At this stage, contrary to the concerns of many experts, we believe that the primary concern for humans should not be the destruction of humanity by “evil” machine intelligence, but rather destruction caused by the stupidity of human decisions leading to the foolishness of machine intelligence.

We argue that the key to advancing towards professional AGI lies in AI's cognitive consistency (including internal and external consistency), being able to self-deconstruct a professionally verifiable reasoning chain, and conducting iterative cognitive evolution stimulated by interactions with other intelligent agents (continuous learning). As AI technology rapidly evolves, its application will significantly expand in various professional domains. However, black-box techniques cannot replace the exploration, accumulation, and iteration of professional knowledge, paradigms, and rules. Only in the way laid out above can AI better serve humanity and bring greater benefits to society.

Wei Tao (韦韬), Vice President and Chief Technology Security Officer at Ant Group, visiting professor at Peking University, member of the Zhejiang Province Association for Science and Technology, and co-founder of notable security academic forum InForSec.

Applications and Exploration in Intel's Responsible AI Work

ZOU Ning and WANG Haining

In the past few years, generative artificial intelligence (AI) has become more powerful – and therefore more capable of doing problematic things in a more convincing and realistic manner. While some have shared concerns about the potential of generative AI to threaten jobs, there is a greater opportunity to responsibly use generative AI to improve people’s efficiency and creativity. We believe that AI should not only prevent harm but also enhance lives.

As part of Intel's Responsible AI work, the company has productized FakeCatcher, a technology that can detect fake videos with a 96% accuracy rate.

Intel’s real-time platform uses FakeCatcher, a detector designed in collaboration with the State University of New York at Binghamton. Using Intel hardware and software, it runs on a server and interfaces through a web-based platform. On the software side, an orchestra of specialist tools form the optimized FakeCatcher architecture. Teams used OpenVino™ to run AI models for face and landmark detection algorithms. Computer vision blocks were optimized with Intel® Integrated Performance Primitives (a multi-threaded software library) and OpenCV (a toolkit for processing real-time images and videos), while inference blocks were optimized with Intel® Deep Learning Boost and with Intel® Advanced Vector Extensions 512, and media blocks were optimized with Intel® Advanced Vector Extensions 2. Teams also leaned on the Open Visual Cloud project to provide an integrated software stack for the Intel® Xeon® Scalable processor family. On the hardware side, the real-time detection platform can run up to 72 different detection streams simultaneously on 3rd Gen Intel® Xeon® Scalable processors.

Most deep learning-based detectors look at raw data to try to find signs of inauthenticity and identify what is wrong with a video. In contrast, FakeCatcher looks for authentic clues in real videos, by assessing what makes us human — subtle “blood flow” in the pixels of a video. When our hearts pump blood, our veins change color. These blood flow signals are collected from all over the face and algorithms translate these signals into spatiotemporal maps. Then, using deep learning, we can instantly detect whether a video is real or fake.

Generative AI can also be used to make 3D experiences more realistic. For example, Intel's CARLA is an open source urban driving simulator developed to support the development, training and validation of autonomous driving systems. Using generative AI, the scenes surrounding the driver would look more realistic and natural, promoting the development of autonomous driving technologies.

GAI can also be used to improve the lives of disabled people. There is a speech synthesis project from Intel, aiming to enable people who have lost their voices to talk again. This technology is used in Intel's I Will Always Be Me digital storybook project in partnership with Dell Technologies, Rolls-Royce and the Motor Neuron Disease (MND) Association. The interactive website allows anyone diagnosed with MND or any disease expected to affect their speaking ability to record their voice to be used on an assistive speech device.

Zou Ning, Senior Director, Technical Policy and Standards at Intel China Ltd.

Wang Haining, Director of AI Technical Policy and Standards at Intel China Ltd.

Acknowledgements

Here, we would like to express our special thanks to all the experts for sharing their valuable insights, which have significantly influenced the quality and depth of this report. Our sincere gratitude goes to Gong Ke and Duan Weiwen for their advice and professional guidance during the planning and writing of the report.

Disclaimer

The views expressed in this article are solely the authors' personal opinions and do not necessarily reflect the position of any institutions. While the report editors and their respective organizations strive for the accuracy, completeness, and reliability of the content, they cannot make any guarantees or promises, and are not responsible for any direct or indirect losses or damages resulting from this report.

Contact

WFEO-CEIT: wfeo-ceit@wfeo.org
Shenzhen Association for Science and Technology: kxbgs@shenzhen.gov.cn

Contributions

Hosting Institutions: WFEO-CEIT, Shenzhen Association for Science and Technology
Report Chief-Editors: WFEO-CEIT Big Data and AI Working Group, Concordia AI