香港科技大學
THE HONG KONG
UNIVERSITY OF SCIENCE
AND TECHNOLOGY

清華大學
Tsinghua University

Developing a Global Framework for AI Governance

Summary Report of the International AI Cooperation and Governance Forum 2023

(Draft)

DISCLAIMER

TABLE OF CONTENTS

Preface 1

Main Forum I 2

Main Forum II 4

Session: AI: Today and Future 6

Session: AI Ethics and Governance 8

Roundtable: Developing A Global Framework for AI Governance 10

Session: AI Industry Development and Governance 11

Session: Frontier AI Safety and Governance 12

Session: AI for Sustainable Development 14

Session: New Paradigm of Artificial Intelligence Empowering Social Science Research 16

Acknowledgments 17

Preface

On December 8-9, 2023, Tsinghua University and The Hong Kong University of Science and Technology (HKUST) jointly organized the International AI Cooperation and Governance Forum 2023. Over two days, the Forum convened more than 50 world-renowned artificial intelligence (AI) experts, scholars, industry leaders, government representatives, and delegates from international organizations. The aim was to delve into the opportunities and challenges presented by cutting-edge technologies such as generative AI and to deliberate on strategies for establishing a comprehensive global governance framework for this new technology.

In recent years, the rapid advancement of generative AI has marked a new era in AI research and development. However, its applications raise numerous security considerations. Under the theme of "Developing a Global Framework for AI Governance," the Forum highlighted the acute sensitivity of the global academic, industrial, and policy communities to international AI governance. It served as a platform for stakeholders to discuss global AI governance issues and future development, fostering basic consensus among the international community on building a global AI governance framework.

The Forum garnered support from organizations worldwide, including the United Nations Development Programme China (UNDP China), UNESCO Multisectoral Regional Office for East Asia, and United Nations University Institute in Macau (UNU Macau). It attracted representatives from government departments in the Mainland, European Union, Singapore, Brazil, South Africa, and Malaysia, as well as academicians from the Chinese Academy of Engineering (CAE) and the Chinese Academy of Sciences (CAS).

Distinguished guests of the Forum included:

- Xinning Lu, Deputy Director of the Liaison Office of the Central People's Government in the HKSAR (CLO)
- Jianming Fang, Deputy Commissioner of China's Foreign Ministry in the HKSAR
- Dong Sun, Secretary for Innovation, Technology and Industry of the HKSAR Government
- Harry Shum, HKUST Council Chairman
- Nancy Ip, President of HKUST
- Hongwei Wang, Tsinghua University Vice President
- Lan Xue, Dean of Schwarzman College and Institute for AI International Governance at Tsinghua University
- YBhg. Datuk Ts. Dr. Mohd Nor Azman Hassan, Deputy Secretary General of the Ministry of Science, Technology and Innovation of Malaysia
- Weiming Wang, Director-General of Educational, Scientific & Technological Affairs at CLO
- Tshilidzi Marwala, Under-Secretary-General of the United Nations and Rector of the United Nations University
- Manuel Innocencio de Lacerda Santos Jr., Consulate General of Brazil in Hong Kong
- Thomas Gnocchi, Ambassador of the European Union office to Hong Kong and Macao
- Mojalefa Mogono, Consulate General of South Africa in Hong Kong & Macau

The report provides a summary of the main sessions and sub-sessions.

Main Forum I

The session featured four prominent speakers: Brad Smith (Microsoft), Wen Gao (Peng Cheng Laboratory), Qionghai Dai (Tsinghua University), and Stephen Cave (University of Cambridge). The session was moderated by Ke Gong (The Chinese Institute of New Generation Artificial Intelligence Development Strategies; Haihe Laboratory of Information Technology Application Innovation).

Brad Smith pointed out that our generation has achieved something unprecedented in human history - we have created machines that can think in place of humans. This calls for us to work together to ensure adherence to some fundamental principles. First, technology should always be under human control. It should serve humanity, and help us to solve some of the biggest problems that affect us all around the world. Secondly, Smith stressed the need to establish a layer at the bottom of technical standards. The legal and regulatory architecture for AI should be based on the technology architecture of AI itself. Thirdly, Smith highlighted the necessity to create computational research of AI. Technological companies should create AI research resources that academic institutes can have access to. This will help academic institutes study new solutions for AI that work in a way that serves the common good. Ultimately, Smith held that all of this requires the continuous building of a responsible AI architecture. Since we are the first generation to create machines that can reason and think, it is our responsibility to make decisions that are in the best interest of all people.

In his address on "Security Issues in the Development of Next Generation AI," Wen Gao emphasized the dual perspective of "AI for Good." From a technological standpoint, he stressed the importance of ensuring that AI meets high standards of efficacy. Simultaneously, from an ethical viewpoint, Gao underscored the imperative of harnessing AI to benefit society positively. Gao highlighted the significant advancements represented by AI 2.0 yet cautioned against the corresponding increase in safety threats. He identified emerging security and privacy concerns within AI systems themselves, posing challenges to the ethical and technological evolution of AI. Gao advocated for systematic prevention and governance of malicious AI usage to address these challenges. This includes enhancing model interpretability, refining algorithm and hardware reliability, and enabling greater control over autonomous decision-making processes. From a technological angle, Gao recommended prioritizing the alignment of natural language model training with Chinese language and values, particularly on domestically driven independent software and hardware platforms. Additionally, he emphasized the critical importance of safeguarding user safety and privacy while leveraging the value of big data, which has become a new factor of production. Gao proposed using technologies like DPI and "Waterproof Fortress Technology" to fortify data security and privacy protections within AI intelligent computing platforms.

During his presentation on "Optoelectronic Intelligent Computing," Qionghai Dai covered the main challenges and advancements in AI technology. He noted the historical progression of AI and highlighted the dilemma posed by the rapid expansion of network scale juxtaposed with sluggish improvements in energy efficiency, thereby limiting computing power. Dai stressed the significance of innovative chip architecture as a breakthrough solution to meet the pressing computational needs of modern AI. Optical computing has emerged as a promising avenue since 2010, offering remarkable improvements in speed, energy efficiency, and data throughput. The development of an All-analog Chip Combining Electronic and Light Computing (ACCEL) has resulted in a staggering million-fold increase in system-level energy efficiency. Recognized by *Nature* for its potential impact, ACCEL is poised to integrate the architecture into daily life sooner than anticipated. The applications of this groundbreaking technology include Unmanned Systems, 5G+Smart City infrastructure, Cloud Computing, and the development of Large language models (LLMs). Dai projected the next phase of AI computing to involve the evolution of general-purpose and AI-specific chips, building upon the mainstream development trajectory. He envisioned novel computing architectures paving the way for a new pathway

in AI advancement.

Stephen Cave delivered a presentation titled “Three Waves of AI Ethics,” discussing the hopes and fears of human-like AI. These include aspirations for ‘immortality’ and concerns about ‘inhumanity’, desires for ‘ease’ contrasted with fears of ‘obsolescence’, hopes for ‘gratification’ countered by concerns about ‘alienation’, and aspirations for ‘dominance’ alongside fears of ‘uprising’. Cave highlighted how these perceptions, though sometimes disconnected from technological reality, can significantly influence AI development, deployment, and regulation. Cave also addressed current affairs, focusing on machine learning and the principles approach. He noted the rapid advancement of machine learning and the emergence of concerns regarding transparency, accountability, responsibility, fairness, and privacy. Despite the proliferation of principles, Cave observed that many still come with limitations, often resembling soft laws with narrow scopes and internal contradictions. Lastly, Cave presented a broader perspective on the rapid technological transformation driven by AI. He warned that such a revolution could spark civil wars, interstate conflicts, or even global warfare. Additionally, he discussed the transformative effects on economic and power dynamics and the externalities and environmental impacts that may arise. Cave emphasized the crucial need for international cooperation and governance to address the ethical challenges posed by AI, stressing that collaboration and governance on a global scale are indispensable.

Main Forum II

The session featured four speakers, including Bo Zhang (Tsinghua University), Dame Wendy Hall (University of Southampton), Yoshua Bengio (Université de Montréal), and Shahbaz Khan (UNESCO Multisectoral Regional Office for East Asia). The session was moderated by QianXiao (Tsinghua University).

Bo Zhang delivered a speech on "Security and Governance in the Era of Generative AI." He pointed out that generative AI, especially multi-modal, can generate human-level images, voices, and videos in the open domain based on text prompts. However, these models lag behind humans in many real-world scenarios, with security issues such as error outputs and malicious uses by exploiting the defects of models, leading to bias, information leakage, fake news, and even unemployment. In this case, there arise two governance problems - model insecurity and user abuse - that require different governance approaches. One is AI alignment, leveraging human-computer interaction to overcome errors and biases; the other is abuse prevention, strengthening relative laws, standards, and regulations. He emphasized that global efforts are essential for governing AI, particularly its users.

Dame Wendy Hall shared her research on "Developing a Global Framework for AI Governance." She reviewed the history of AI development, emphasizing its periods of prosperity and challenges. With the advent of generative AI, powered by significant computing and data resources, she discussed the opportunities for humanity in health services, intelligent cities and transport, and climate change, as well as the risks that AI brings. As Stephen Hawking said, the development of full artificial intelligence could spell the end of the human race. Hall stressed the need for social responsibility in building AI systems, considering ethical principles. While tech companies currently lead AI development, their primary agenda is profit rather than AI for good, necessitating a sensible global framework to balance growth and manage risks. Hall called for the involvement of tech companies, governments, and research institutes in this process.

Yoshua Bengio shared insights on "Democratic Governance to Manage AI Risks," highlighting that intelligence grants power over other species on Earth. As we imbue machines with intelligence, this power increases, raising concerns about potential abuse. The looming question is what will happen when machines surpass human intelligence. Bengio emphasized the responsibility of human society to contemplate how to avoid the harm caused by these powerful machines, identifying two key aspects for consideration. First, there is a need to address the challenge of alignment and control, ensuring that machines are used in ways that don't harm humans. Second, preventive measures must be implemented to curb malicious uses, involving regulations and preparedness for instances where a dangerous situation arises despite due diligence. Bengio expressed the current inadequacy of countries in establishing rules and standards for monitoring AI development. To make informed decisions in managing AI risks, Bengio advocated for robust, independent, and democratic oversight. This oversight should involve national regulators and incorporate civil society, independent academics, and the international community. He presented three recommendations:

1.Implement national regulations and international treaties, restricting access and frontier development to licensed organizations and registered models. This framework should include audits and the prohibition of unproven models.

2.Allocate significant research investment in AI safety to enhance understanding of risks and guide regulatory efforts.

3.Develop a contingency plan (Plan B) to address potential AI-driven threats arising from misuse or loss of control.

In his address, Shahbaz Khan highlighted that generative AI signifies a technological paradigm shift. These technologies have showcased impressive capabilities, from crafting personalized educational

content to developing predictive models for mitigating climate change. However, they also bring forth substantial challenges. Ensuring that AI is a tool for enhancing our global society is paramount. This notion involves respecting cultural diversity and preventing AI from exacerbating technological disparities within and between nations. Khan emphasized the significance of UNESCO's Recommendation on the Ethics of Artificial Intelligence, which offers a vital framework. The recommendation advocates for a balanced approach that maximizes the benefits of AI while concurrently addressing and mitigating its risks. The ultimate goal is to foster fair and inclusive outcomes in deploying AI.

Session: AI: Today and Future

Innovations in AI are continuously shaping the future of humanity across various industries, driving emerging technologies like big data, robotics, and IoT. Generative AI, in particular, has expanded the possibilities and popularity of AI, solidifying its role as a key driver of progress.

The "AI: Today and Future" session featured four distinguished speakers: Jiaya Jia (The Chinese University of Hong Kong and SmartMore), Xiaochuan Wang (Baichuan AI), Xing Xie (Microsoft Research Asia), and Laurence Devillers (Sorbonne University). The session was chaired by Maosong Sun (Tsinghua University).

Jiaya Jia first discussed the "IndustryGPT: World's First Large Language-Vision Model for Industrial Automation," where he spoke about the current uses of large models in the industry. He noted that no large language model currently can cater to the high-end manufacturing industry. SmartMore Corporation specializes in providing comprehensive products and solutions for smart manufacturing and digital innovation, and they have created IndustryGPT V1.0 for industrial applications. To develop IndustryGPT, developers require a large amount of industrial-related knowledge, which they refine by using the large model. To further improve IndustryGPT, it's essential to have interactions between models and humans, as well as between software and hardware.

Xiaochuan Wang delivered a speech on "The Practice and Outlook of Baichuan Intelligence in the Era of Large Models," stressing that when machines master language, the era of general AI will have arrived. The rapid technological revolution has brought forth the era of generative AI, and large models are illuminating intelligence across various fields in society. Wang introduced a solid tech stack consisting of foundation models and search capabilities. This means that the foundation model R&D is based on past search engine practices (data, algorithms, and computing power), and the search engine technology supplies the foundation model (hallucination, timeliness, and domain knowledge). Wang also reviewed the history of Baichuan Intelligence and shared its practice in developing open-source models.

Xing Xie shared his research on "Societal AI: Tackling AI Challenges with Social Science Insights." In his presentation, he reviewed the evolution of social sciences and AI in society, highlighting that both these fields have evolved to better understand human behavior and intelligence. He also emphasized that the convergence of social science and AI can equip us to better understand and prepare for the future. Xie suggested two recommendations for leveraging social science insights to tackle AI challenges. The first is to evaluate general-purpose AI with psychometrics, which provides reliable measurements of mental capacities and psychological attributes. He introduced a framework for evaluating AI systems with psychometrics, which involves the construction of identification, measurement, and validation. He stressed the need to redefine the terms "person" and "population," identify the sensitivity and variability of AI, and recognize AI and humans. The second recommendation is the value compass, which is a new paradigm for AI value alignment. The traditional process of basic value alignment covers behavior collection and value identification. Xie's new paradigm for AI value alignment includes the following:

1) AI basic values benchmarking: a comprehensive dataset capturing AI behaviors and their basic values.
2) Value judgment modeling: training the model to discern basic values flexibly and evaluate alignment based on interactions among agents.
3) Automatic value evaluation: producing generative evaluation questions tailored to each model and moving to dynamic evaluation.
4) Value verification and behavior prediction: confirming AI values with clarity and predicting future AI risk behaviors.

5) Efficient and adaptive alignment: crafting algorithms based on Schwartz's theory, accessing the alignment of LLMs, and achieving alignment with pluralistic values.

Laurence Devillers presented "Generative AI and Affective Computing in HM Spoken Interaction: Today and Future Issues." During the presentation, Devillers pointed out that since the 19th century, the analysis of technological advancements has tended to be viewed through the lens of binary oppositions between progress and disaster. This approach is also evident in companies seeking to create a "general artificial intelligence" that could rival or surpass human intelligence. By promoting such ideas, developers of AI systems like ChatGPT are simultaneously fueling both hopes and concerns. Generative AI can produce multiple outputs from multimodal inputs, and its impact on society and the economy is likely to be significant in various areas of application. However, generative AI development raises ethical, psychological, economic, social, cultural, and governance issues. These issues arise partly due to the lack of reliable evaluation methods for model design and the limited and flawed nature of available data. Devillers provided 12 recommendations on generative AI governance:
1) Creating a sovereign research and training entity for "AI, science and society."
2) Accelerating the adoption by economic stakeholders.
3) Sharing practices in the use of generative AI systems.
4) Facilitating regulation of watermarks.
5) Using generative AI systems in education.
6) Providing open access to foundation models.
7) Treating foundation models introduced into the market and generative AI systems as high-risk AI systems.
8) Creating a chain of responsibility.
9) GDPR and generative AI systems.
10) Processing of collected data.
11) Understanding copyrights in relation to generative AI.
12) Emphasizing the environmental impact of generative AI.

Session: AI Ethics and Governance

In recent years, the rapid growth of AI technologies has sparked widespread discussions about their transformative potential in addressing complex issues across various domains. Apart from boosting workplace productivity, there is a growing recognition of AI's impact on medicine, healthcare, sustainable development, the environment, and climate change. With this surge in AI capabilities, there is a growing need for both comprehensive governance and ethical frameworks to guide its responsible deployment. The myriad global complexities of AI ethics and governance thus demand a holistic approach, as piecemeal solutions often prove inadequate in addressing the entirety of these intricate concerns.

The AI Ethics and Governance session featured six keynote speakers who explored the current opportunities and challenges of AI ethics and governance, as well as future actions in developing equal and responsible AI. The speakers included Chloé Bakalar (Meta), Joaquin Quiñonero Candela (LinkedIn), Haitian Lu (The Hong Kong Polytechnic University), Shannon Vallor (University of Edinburgh), Christoph Lütge (Technical University of Munich), and Wendell Wallach (Carnegie Council for Ethics in International Affairs). The session was chaired by Pascale Fung (HKUST).

Chloé Bakalar and Joaquin Quiñonero Candela presented their research on "Practical AI Fairness: Math, Ethics & Politics," emphasizing that there is a paradox among the math, ethics, and politics disciplines due to there being multiple, incompatible definitions of fairness, which poses a challenge to disentangle and operationalize AI fairness. Besides, AI fairness depends on the specifics and context of the product in which AI is deployed. Given this concern, AI fairness requires disentangling fairness contradictions, and a potential approach is to separate equal AI treatment from product equity: AI systems must guarantee equal treatment and equitable outcomes should be driven by product strategy. Equal AI treatment is measured quantitatively via predictive parity and is math-friendly; product equity has to be grounded in a deep understanding of patterns and barriers of various communities. AI fairness is not a zero-sum game; in other words, it is a false choice between equal AI treatment or product equity. Since AI creates new urgency for old problems, philosophers and political scientists must be involved in the discussions.

Haitian Lu presented research on "AI and the Future of Legal Services: Ethical Implications." His presentation highlighted how AI empowers the legal service market, ethical considerations in AI adoption, and potential solutions for future AI ethics, dispelling the misconception that lawyers resist technology adoption. In fact, recent advancements in machine understanding and the generation of text have unveiled groundbreaking possibilities. LLMs are highly applicable in the legal field, as LLMs excel in understanding (through natural language processing) and generating (using natural language generation) legal content. However, significant challenges on AI ethics have raised ethical considerations for lawyers in AI adoption. AI ethics is a field where conventional wisdom meets new problems, requiring a holistic approach covering law, technology, governance, finance, and education and inter-disciplinary dialogues, particularly dialogues between social science scholars and computer scientists. Six interconnected tasks were recommended for addressing ethical challenges in AI: investigation of high-level principles; survey of the societal values; AI ethical incidence case databases; AI ethics classifying frameworks; establishment of a normative baseline; and ethics-embedded test cases.

Shannon Vallor delivered a presentation on "Reconceiving AI Governance as Social Care," highlighting the ethical implications of responsibility gaps in AI governance that lead to severe moral consequences, eroding social trust. To address this, AI governance should transition from a control-oriented model that only requires confident prediction of AI capabilities and effects to one rooted in social care, prioritizing the protection and well-being of vulnerable communities and groups over centralized

regulatory mechanisms. The proper and necessary function of governance is not control. Instead, it is the protection and care of the moral ecology of a group or community. This shift requires a moral attitude that listens to and responds to the concerns of vulnerable communities, emphasizing the importance of responsible AI governance as a form of social answerability. Building trustworthy AI must first be allocated to those who are jeopardized voices, listen to those communities, and respond to their articulations of the danger.

Christoph Lütge presented his research on "Current Opportunities and Challenges for AI Governance and Ethics," beginning with a review of the G7 Outcomes and Implications for Global AI Governance, a significant outcome this year, with the emphasis on the statement that while "the common vision and goal of trustworthy Al may vary, the governance of the digital economy should continue to be updated in line with common values." It is undeniable that AI presents enormous opportunities for humankind, especially in healthcare. Still, the current AI system remains potentially harmful to ethics, including algorithms, data privacy, reliability, bias, fairness, and transparency. Concerning governing the risks of Al while benefiting from the opportunities, there are three approaches for building a trustworthy Al: lawful, ethical, and robust. Various forms of governance are presented in a lawful AI model, namely, a top-down form with both hard law and soft law tools, a bottom-up form of self-regulation in all its variants, and a "middle-out" form incorporating both elements of top-down legal framing and bottom-up empowerment of individual actors. To ensure AI is for everyone, a "middle-out" approach is suggested.

Lastly, Wendell Wallach presented research on "A Global AI Observatory and Soft Law Functions in the International Governance of AI." In the AI field, soft law functions refer to high-level tasks or objectives that support the responsible development and use of Al models, which play roles in assessing opportunities, risks, and impact, international cooperation, expectation setting and harmonization, policy design support, and evaluating and monitoring AI systems. In view of more serious regulation of AI in the future, it becomes necessary to create a Global AI Observatory (GAIO) to facilitate communication, cooperation, and coordination among the many initiatives entering the AI governance space. The GAIO's responsibilities should cover various tasks, ranging from producing annual reports to maintaining critical registries. Moreover, the international community must take immediate action to formulate a durable governance framework that enables effective ethical and legal oversight of AI. The proposed framework consists of five symbiotic components:

1) Neutral Technical Organization: This entity will continuously evaluate the global acceptance levels of legal frameworks, best practices, and standards related to AI;
2) Normative Governance Capability: With limited enforcement powers, this part aims to encourage adherence to global standards for the ethical and responsible use of AI and its associated technologies;
3) Conformity and Certification Tools: These tools will evaluate and certify compliance with established standards;
4) AI-Governance-Supporting Tools: These are dedicated tools that aid in dealing with data associated with decision-making, validating and auditing existing systems, and addressing risks as needed; and
5) Global AI Observatory: As a bridge to narrow the gap in understanding between scientists and policymakers, this observatory will fulfill the functions defined below not already covered by other institutions.

Roundtable: Developing A Global Framework for AI Governance

The roundtable discussion on "Developing A Global Framework for AI Governance" brought together six distinguished panelists: Pascale Fung (HKUST), Zheng Liang (Tsinghua University), Rostam J. Neuwirth (University of Macau), Serge Stinckwich (UNU Macau), and Jianrong Tan (Zhejiang University). The session was chaired by Liesbeth Venema (Nature Machine Intelligence).

Pascale Fung emphasized the necessity of grounding AI governance in global and multinational discussions. Recognizing the influence of different cultural contexts on views about AI governance, Fung stressed the importance of open-mindedness among technical experts and policymakers to facilitate global conversations and understanding. Fung underscored the humanistic dimension in advancing AI governance, urging technical professionals to delve into human interests and values. Fung asserted that AI governance systems should not be developed in isolation but instead consider global dynamics. Policymakers, in turn, should familiarize themselves with technology, especially in the context of frontier AI technologies.

Zheng Liang brought attention to two pivotal concepts in the establishment of a global AI governance framework. The first, "value," underscores the need for AI development to develop a nuanced understanding that extends beyond technical considerations. The process of infusing human values into AI systems requires a global dialogue where individuals share and negotiate diverse cultural perspectives to reach a consensus. Defining these values is a crucial process, as diverse cultures contribute to different perspectives. Second, Zheng emphasized the "Global South," stressing the underrepresentation of these voices in global AI governance discussions. He called for concerted efforts from every nation to build a unified framework, recognizing the importance of capacity-building and addressing digital gaps in the Global South.

Rostam J. Neuwirth indicated that a global framework for AI governance is based on a common ground rather than a single legal system. People need to rethink the existing international legal system, which was primarily developed post-World War II, and whether this existing legal system is sufficient for addressing AI governance in the real world. AI, Augmented Reality, and Immersive Technologies are constructing a more coherent and multisensory reflection of human beings – the stimulus of one sense triggers a reaction in another. However, at least in the law field, this multisensory reality has not been fully considered. AI governance requires establishing a consistent, coherent, multisensory global AI governance system.

Serge Stinckwich highlighted the significant potential of AI in contributing to the Sustainable Development Goals (SDGs) but also raised concerns about the challenges it presents. Specifically, he pointed to the emergence of generative AI as posing new threats to the SDGs. Stinckwich observed that technology companies often develop tools without clearly understanding their implications. This shortcoming leads to a reactive approach in addressing consequences post-deployment. He emphasized the crucial role of the United Nations in spearheading governance efforts related to generative AI. He advocated for a deeper understanding of technology impacts. Stinckwich proposed two recommendations: first, integrating governance considerations into the design phase of AI, and second, adopting a more holistic social-technical system approach.

Jianrong Tan stated that prioritizing the governance of AI is essential to unlock its benefits fully. It is crucial for people to have a clear understanding of what aspects of AI should be governed and to leverage technology in governing AI effectively while minimizing potential harm. Given that humans are the creators and developers of machines, the rapid advancement of technology poses a significant challenge: how to govern machines when their intelligence surpasses that of humans. Tan underscored the need for technical experts to contemplate and address these risks during AI development by implementing appropriate measures. In essence, AI is propelling humanity towards a cognitive revolution, with the overarching goal of harnessing AI for the greater good.

Session: AI Industry Development and Governance

Modern AI is a resource-intensive field that relies on extensive datasets, skilled data scientists, and substantial computing power. Typically, these resources are concentrated within a few large companies, prompting concerns about the potential worrisome concentration of power in AI development and its future trajectory. Therefore, the active involvement of academic institutions in the AI field is of paramount importance, given that these institutions often prioritize public interest and the advancement of knowledge in AI research and development.

The session on AI Industry Development and Governance showcased four keynote speakers who shared their insights on fostering collaboration between academia and industry. The keynote speakers were Lei Fang (DataCanvas), Bifei Mao (Huawei), Zhiqing Yin (QI-ANXIN Group), and Jiang Liu (Zhipu AI). The session was chaired by Lei Chen (HKUST GZ).

Lei Fang delivered a speech on "The AI Native Enterprise Era," emphasizing that enterprises will be driven by intelligence in the future rather than by digital transformation and software automation. The era of large models necessitates upgrades from different perspectives. From a technology perspective, it requires a complete infrastructure upgrade in terms of computing power, data, and foundational software. From a value perspective, it requires the alignment of human values in AI.

Bifei Mao shared her research on "Choices of a Practical AI Governance Framework," highlighting that the widespread application of AI requires a gradual evaluation of its potential societal influence. AI governance is a long-term exploration requiring continuous technological innovation and communication among all stakeholders. A practical AI governance framework involving the public, regulatory bodies, certification entities, users, and providers, with effective coordination and collaboration, can mitigate challenges and risks associated with AI.

Zhiqing Yin discussed "Exploration and Practice of Using AI to Create Industry-level Productivity in Network Safety," underscoring that cybersecurity faces productivity shortages and AI presents an effective solution to elevate large models to an industrial level, enhancing productivity. The three core elements of industrial-level cybersecurity large models include high-quality safety data, multidisciplinary professional teams, and diverse scenarios.

Jiang Liu provided insights on "The Safety/Alignment of Large Language Models in Practice," noting that the more intelligent AI becomes, the higher the potential for harm can be. Current safety concerns for large models encompass cognitive, social, and political dimensions. It's crucial to ensure that large models align with human values and foster increased alignment among humans. International cooperation and governance will play a pivotal role in promoting the safety and alignment of large models. The advancement of frontier AI relies heavily on the contributions of the scientific community and AI developers, particularly in ensuring its safety. Equally vital are national policymakers and the international community, tasked with fostering the healthy development of AI while mitigating associated risks.

Session: Frontier AI Safety and Governance

The Frontier AI Safety and Governance Session featured thirteen distinguished speakers, including Bo Zhang (Tsinghua University), Bowen Zhou (Tsinghua University), Jimmy Ba (xAI), Michael Sellitto (Anthropic), Sean S. ÓhÉigeartaigh (University of Cambridge), Xing Xie (Microsoft Research Asia), Jie Fu (HKUST), Nicolas Mialhe (The Future Society), Michael Frank (Center for Strategic and International Studies), Qiqi Gao (East China University of Political Science and Law), Wan Sie Lee (Infocomm Media Development Authority), Qi Chen (Tsinghua University), and Angela Zhang (The University of Hong Kong). The session was chaired by Brian Tse and Kwan Yee Ng (Concordia AI).

Bo Zhang highlighted the rapid ascent of generative AI and large models, igniting the dawn of artificial general intelligence. While these advancements empower various industries, they also pose a series of challenges. Large models, exemplified by GPT, demonstrate emergent behavior and unexpected capabilities, yet they can also manifest hallucinations and lack robustness and self-awareness. Strengthening the governance of large models demands effective technical measures alongside addressing governance issues to prevent misuse and abuse. Moreover, there's a call to pursue the development of a "third generation of AI," fostering new interpretable and robust AI theories and methods and addressing fundamental AI safety concerns.

The first discussion was prompted by the question, "What can the scientific community and AI developers do to support frontier AI safety and governance?"

Jimmy Ba outlined the significance of foresight, which involves understanding the potential progression of AI models and insight into how LLMs operate. He stressed the importance of oversight in AI development. Additionally, he proposed guiding students to critically assess AI-generated content instead of limiting their exposure to generative AI. Finally, he underscored the necessity for collaboration between generalists and technical experts to foster AI development and governance.

Bowen Zhou delivered a speech titled "Supporting the Governance of Foundation Models in Full Life Cycle by the Scientific Community." Zhou emphasized the importance of governing the complete AI model lifecycle to address uncertainties and risks, including existential threats. He proposed that scientists could act as coordinators between stakeholders in AI governance, actively govern AI across its entire lifecycle, and allocate greater resources toward building trustworthy AI.

Michael Sellitto introduced Anthropic's AI Safety Levels (ASL) framework, which establishes increasing security standards corresponding to rising risks to promote responsible AI development. He highlighted that the responsibility for ensuring safety grows with the advancement of capabilities.

Sean S. ÓhÉigeartaigh raised concerns about the practice of open-sourcing AI, noting its potential benefits alongside associated risks. He emphasized that not all frontier models are suitable for open sourcing. Furthermore, he argued that neither a fully open nor fully closed approach ensures safety, highlighting the importance of oversight mechanisms.

During the panel discussion, Xing Xie advocated for interdisciplinary research to develop more reliable evaluations of AI. Bowen Zhou emphasized the challenge of aligning AI with human values. Regarding AI safety funding, Zhou also predicted that training AI models would require less computing power than aligning them in the future. In conclusion, speakers provided brief recommendations for the scientific community and AI developers to support frontier AI safety and governance. Suggestions included establishing more national initiatives similar to the UK AI Safety Institute to educate policymakers and the public on frontier AI, organizing more inclusive conferences to promote

international engagement and mutual understanding, conducting additional model evaluations, and implementing programs to encourage interdisciplinary research in AI safety and governance.

The second discussion was prompted by the question, "How can policymakers and the international community work together to improve frontier AI safety and governance?"

Nicolas Mialhe discussed "Governing the rise of general-purpose AI: Taking stock of ongoing international cooperation efforts and possible pathways." He introduced a "functionalism" approach to global AI governance, suggesting that initial alignment on goals such as peace and prosperity should precede the coordination of strategies and standards. Mialhe argued that international cooperation is essential to address AI's rapid pace and impacts, emphasizing the need for cooperative efforts among different mindsets and structures.

Michael Frank, Senior Fellow at the CSIS Wadhwani, addressed "International Approaches and Possibilities for Cooperation on AI Governance." He examined the variances and trade-offs in AI governance models, contrasting the EU's legislation-based risk tiering with the US's executive order approach. Frank concluded that despite these differences, international cooperation remains both possible and essential, driven by shared risks and incentives across nations.

In the discussion on "Consensus Governance of LLMs from the Perspective of Intersubjectivity," Qiqi Gao introduced potential "consensus governance" systems for overseeing LLMs at corporate, national, and global levels. He argued that technical alignment alone is insufficient and emphasized the need for multi-stakeholder governance, including third-party auditing. Gao warned that escalating societal risks may necessitate the expansion of participatory governance and oversight frameworks.

In her presentation on "Evaluation and Testing for Generative AI," Wan Sie Lee outlined Singapore's objectives with the AI Verify governance initiative to ensure trustworthy AI that meets global standards and addresses stakeholders' needs. She delved into the challenges associated with capabilities for risk-proportional oversight and testing collaboration across various AI models.

During the panel discussion, Angela Zhang highlighted how geopolitical tensions are eroding AI safety and governance efforts, resulting in countries underregulating AI. She emphasized the critical need for more dialogue and cooperation in addressing these challenges. Qiqi Gao emphasized the urgency for the international community to establish more explicit definitions of generative AI to enhance governance measures, describing the situation as an "emergency" amidst AI's rapid development. Qi Chen stressed the importance of continued collaboration among epistemic communities worldwide, advocating for a bottom-up approach. Overall, there was a consensus that AI poses challenges beyond national borders, necessitating coordinated global governance based on shared priorities. The participants agreed on the importance of international governance frameworks in categorizing AI risks and developing corresponding response plans. However, some cautioned against overly relying on institutional analogies, advocating for a balanced approach.

Session: AI for Sustainable Development

A Nature article from 2020 revealed that while nearly 80% of the SDGs can benefit from AI, 35% may face negative impacts. This dual-edged nature underscores the importance of a balanced approach to AI deployment for sustainable development. The more we rely on AI for sustainable development, the greater care we must take to ensure that it is harnessed for good.

The AI for Sustainable Development Session featured two keynote speakers and six panelists who provided insights on minimizing AI's risks while maximizing its potential to help tackle global challenges and advance SDGs. Keynote speakers included Yike Guo (HKUST) and Xufeng Zhu (Tsinghua University). The session was chaired by Wei Zhang (UNDP China).

Yike Guo discussed "The Human Use of Human Being: University Education in the AI Era," emphasizing the transformative impact of AI on higher education, which can create a new world of learning. The recent progress of generative AI has enabled education innovation, redefining our conventional understanding of "knowledge," "learning," and "teaching." However, in a world where AI-generated content is becoming indistinguishable from human output, the crux of education should pivot to understanding, critiquing, and enhancing these outputs rather than just competing with them. A mindset shift is required from both educators and learners to adapt to this new transformation: from a humanistic perspective, they should focus on human values and capabilities to differentiate humans from machines; from a cognitive perspective, they should emphasize cognitive capacities and knowledge practices; and from a social perspective, they should consider the distribution of agency between individuals and collectives and the impact of AI on society. Suggestions for building best practices in AI-enabled higher education include customized learning experiences for personalized education with AI-driven content adaptation; collaborative learning with AI-enabled peer interaction and group problem-solving; new assessment methods to apply AI for dynamic, adaptive testing of understanding; the use of AI to monitor learning patterns for tailored assessments; automated essay scoring systems and feedback tools that provide useful feedback on student work; and the development of pilot projects for new pedagogical innovations.

Xufeng Zhu presented his research on "Path to Carbon Neutrality through AI," underscoring the significant potential of AI to contribute to carbon neutrality. He emphasized that the development of AIGC (Artificial Intelligence and Green Computing) can significantly enhance efficiency in carbon neutrality initiatives. AI plays a pivotal role throughout the entire life cycle of carbon-neutral actions. Notably, AI technologies are applied in cutting-edge projects such as Digital Earth, an information system that describes, processes, and analyzes the Earth's environment and space activities, and Green City Watch, a program focusing on urban forest management. Instrumental in these initiatives, AI algorithms collect data from multiple sources to determine carbon footprints. Additionally, AI empowers smart grid solutions, contributing to the balance of electricity production and consumption while maintaining grid stability. In China, substantial efforts have been undertaken to address climate change, aiming to peak carbon emissions by 2030 and achieve carbon neutrality by 2060. The rapid development of AI in China has facilitated the deployment of AI-enabled solutions to address challenges related to population aging, promote the development of machine intelligence in an environmentally friendly manner, and even reduce carbon emissions to levels lower than those associated with human labor. However, it is crucial to acknowledge that AI, in its utilization process, can generate a carbon footprint, potentially impacting carbon neutrality efforts. Consequently, there is a growing need for developing and implementing "green AI" solutions that mitigate adverse environmental effects and contribute positively to carbon reduction initiatives.

As per the International Telecommunication Union (ITU), the UN has consistently undertaken comprehensive system-wide efforts, fostering collaboration both within and between organizations. In

the previous year, 40 UN entities initiated 281 activities dedicated to Artificial Intelligence for Sustainable Development Goals (AI4SDGs). This underscores a collective commitment towards leveraging AI to achieve the SDGs. During the panel discussion, representatives from six UN agencies shared insights into their respective endeavors in the policy space. The distinguished speakers included Fengchun Miao (UNESCO), Serge Stinckwich (UNU Macau), Zhongxin Chen (Food and Agricultural Organization, FAO), Janine Berg (International Labour Organization, ILO), Alexandra Hakansson Schmidt (Regional Office for Asia and the Pacific of the United Nations Entity for Gender Equality and the Empowerment of Women, UN Women Asia-Pacific Regional Office), and Sameer Pujari (World Health Organization, WHO).

Fengchun Miao shared UNESCO's groundbreaking global guidance on generative AI in education, titled "Guidance for Generative AI in Education and Research." He highlighted the critical need for educational institutions to validate GenAI systems based on their ethical and pedagogical appropriateness for education. Miao urged adopting a human-agent and age-appropriate approach in this validation process.

Serge Stinckwich introduced UNU Macau's research, focusing on leveraging computational collective intelligence for sustainable development. This innovative approach uses a human-centered methodology to build collective intelligence, promoting collaboration between humans and machines for a more sustainable future for all.

Zhongxin Chen provided insights on how AI could be a powerful tool to revolutionize the agricultural sector by enhancing efficiency, precision, and sustainability and called on multi-stakeholders to design and leverage this digital tool with users in mind.

Janine Berg shared ILO's recent study on "Generative AI and jobs: A global analysis of potential effects on job quantity and quality." The study predicts that the overwhelming impact of generative AI will be to augment occupations rather than automate them. Berg underscored the need for proactive policies focusing on job quality, ensuring fair transitions, and being based on dialogue and adequate regulation.

Alexandra Hakansson Schmidt expressed the view that AI offers innovative solutions to address bias and cyberbullying in online engagement. However, she highlighted the inherent bias in AI systems, emphasizing the potential pitfalls if left unaddressed. This bias may have serious implications for gender equality and peace and security.

Lastly, Sameer Pujari emphasized the evident potential of AI to advance the development of medicines, support health systems management, and enable clinical professionals to improve patient care and perform complex medical diagnoses. However, Pujari stressed that the risks and challenges in AI applications cannot be ignored. It requires careful regulatory considerations to ensure that AI is used safely, effectively, and ethically in healthcare.

Session: New Paradigm of Artificial Intelligence Empowering Social Science Research

In today's landscape, intelligent solutions driven by the approaches, methods, or techniques cutting-edge technologies employ permeate every field. Recent trends indicate that AI dynamics not only guide but also shape the trajectory of research focus within the humanities and social sciences.

The session on New Paradigm of Artificial Intelligence Empowering Social Science Research featured six speakers, including Ke Gong (The Chinese Institute of New Generation Artificial Intelligence Development Strategies; Haihe Laboratory of Information Technology Application Innovation), Yongheng Yang (Tsinghua University), Gang Liu (Nankai University), De Kai (The Hong Kong University of Science and Technology), Hualing Fu (The University of Hong Kong), and Weijie Sun (DP Technology). The session was chaired by Ke Gong and Gang Liu.

Ke Gong emphasized in his speech that AI, as a revolutionary technology, plays a crucial role in social governance and services. AI also presents opportunities to overcome research limitations and enhance social science research. The primary focus in leveraging AI for social science research is effectively managing AI risks.

Yongheng Yang pointed out that modern society is swiftly entering the digital age, with digital technologies, particularly AI, transforming our work and life. This transformation brings new opportunities for theoretical systems and knowledge creation in three key aspects:
1.The algorithmic thinking of AI provides guidelines for new paradigms and methodologies in social science research.
2.AI promotes arts and science, giving rise to new interdisciplinary subjects.
3.AI, along with other digital technologies, creates a digital world paralleled with the real world.

Gang Liu emphasized that with the acceleration of economic globalization and the new technological revolution, science-driven economic and social development has become complex and uncertain, posing challenges to social science research. Current research methods fall short in meeting these new challenges. A solution is the specialization of general technologies - establishing a platform and system for social science research driven by knowledge and data. This platform marks a new direction for applying AI in social science research.

De Kai shared his research on "Misframing AI Governance," underscoring the need to reduce the deployment of AI that can exacerbate social instability and accelerate technology development to minimize negative impacts. He indicated that the core factor of misframing AI governance is the long-term instability of human civilization. Therefore, consideration for human civilization should be incorporated into future AI development.

Hualing Fu delivered a speech on "Would AI Bring a Paradigmatic Change in Legal Research." He noted that the development and popularization of AI present new challenges to social laws and lawyers, including:
1.For legal education in universities, teachers should consider how to cultivate students with digital literacy.
2.For legal judgment, there is a controversy over whether AI can replace judges in making conclusions.
3.For legal research, the bias between legal judgment and AI prediction requires further exploration.

Weijie Sun explored "AI for Science," highlighting that the application of AI for social science is rooted in interdisciplinary research. The governance of AI technology and large models identifies new research needs in sociology and ethics. Simultaneously, research in psychology, economics, and complex sociology demands AI technologies. AI is reshaping the landscape of social science research and propelling it into new frontiers.

Acknowledgments

The report has greatly benefited from the insight and expertise of the organizers and speakers.

Special appreciation goes to the **Organizing Committee of the Forum**:
- Xiqin Wang, President of Tsinghua University
- Nancy Ip, President of HKUST
- Lionel Ni, President of HKUST (GZ)
- Bin Yang, Vice President of Tsinghua University
- Yike Guo, Provost of HKUST
- Hongwei Wang, Vice President of Tsinghua University
- Tim Cheng, Vice-President for Research and Development of HKUST
- Yang Wang, Vice-President for Institutional Advancement of HKUST
- Yongheng Yang, Director of the Office of Arts, Humanities and Social Sciences Administration of Tsinghua University
- Zheng Liang, Vice Dean of the Institute for AI International Governance of Tsinghua University
- Andrew Cohen, Director of Jockey Club Institute for Advanced Study of HKUST
- Qian Xiao, Vice Dean of the Institute for AI International Governance of Tsinghua University
- Kellee Tsai, Dean of the School of Humanities & Social Science and Associate Director of the Centre for Artificial Intelligence Research of HKUST
- Junqun Lu, Secretary General of the Institute for AI International Governance of Tsinghua University

Special gratitude is also due to the **Academic Committee of the Forum**:
- Harry Shum, Council Chairman of HKUST
- Lan Xue, Dean of the Institute for AI International Governance of Tsinghua University
- Yike Guo, Provost of HKUST
- Hui Xiong, Associate Vice-President for Knowledge Transfer of HKUST (GZ)
- Yaqin Zhang, Dean of the Institute for AI Industry Research at Tsinghua University
- Maosong Sun, Executive Assistant Dean of the Institute for Artificial Intelligence of Tsinghua University
- Pascale Fung, Chair Professor of the Department of Electronic and Computer Engineering and Director of the Center for Artificial Intelligence Research of HKUST
- Jun Su, Director of Think Tank Research Center at Tsinghua University
- Xiaofang Zhou, Head of the Department of Computer Science & Engineering of HKUST
- Xufeng Zhu, Dean of the School of Public Policy and Management and Executive Director of the Institute for Sustainable Development Goals of Tsinghua University
- Lei Chen, Dean of the Information Hub of HKUST (GZ)
- Bowen Zhou, Chair Professor of the Department of Electrical Engineering of Tsinghua University
- Bert Shi, Director of the Center for Aging Science of HKUST
- De Kai, Professor of the Department of Computer Science & Engineering of HKUST

Finally, and most importantly, a debt of gratitude is owed to all distinguished speakers who shared their insightful contributions and to all supporting partners of the Forum.

CONTACT

Author
Chang Liu, Institute for AI International Governance of Tsinghua University,
liuchang3010@tsinghua.edu.cn