Reflections on Further Deepening International Cooperation in the Development of Artificial Intelligence

GAO Wen

Abstract:

The cutting-edge artificial intelligence technology represented by large models has become a key field for new industries and new business forms around the world. However, the rapid development of artificial intelligence technology has also brought various security problems and risks. The core of dealing with risks includes raising the moral awareness of science and technology for good, having sufficient laws and regulations to restrict the evil exploitation, and having sufficient technical means to prevent the evil exploitation. Therefore, it is necessary to carry out relevant international cooperation to achieve cross-broader governance. Pengcheng Laboratory launched the WCNC World Computing Net Consortium, built a multi-language cooperation platform, and actively participated in and prepared relevant international big science programs.

How to Ensure Safe and Trustworthy AI Application Systems for Good

WU Jiangxin

Abstract:
Artificial intelligence(AI) technology is a double-edged sword, with widespread security issues at multiple levels. Research shows that all technological systems have unintended side effects abeyond their designed functions, and uncertain security threats are a major challenge that current cybersecurity paradigms struggle to address. Endogenous Safety & Security（ESS）architecture is an effective solution to these challenges. For AI systems, their inherent 'black box' effect leads to unavoidable internal flaws, referred to as the Three Nons' genetic defects (Mechanism issues) in AI models, which trigger unique ESS problems. Studies indicate that AI models must have necessary diversity across multiple dimensions, which enhances the discriminative capacity of the ESS architecture. This diversity forms the foundational conditions for achieving AI alignment and reaching the axiom of relative correctness. By designing heterogeneous structures and data, and implementing a fusion-decision mechanism to generate system outputs, an AI system with ESS architecture can effectively identify and address generalized security events caused by model differences. The network endogenous security paradigm should be the core direction of AI trustworthy application system theory development.

Toward Culturally Inclusive NLP

Alice OH
Professor
School of Computing, KAIST

Abstract:
Cultural context plays an important role in generating and understanding human language. Tasks that involve subjective judgment such as hate speech and social stereotype detection are strongly influenced by the cultural context. In this talk, I will present recent research on revealing the culturally-insensitive treatment of hate speech, building an English hate speech benchmark datasets that are more inclusive of diverse cultures, and adapting BBQ, an English social stereotype dataset to Korean.

I will also introduce our latest work on building a multilingual, multicultural dataset of commonsense knowledge and share how the state-of-the-art LLMs perform on various languages and cultures. By presenting these recent efforts to explicitly reveal how the current LLMs are heavily skewed toward the major cultures in their cultural competence, I hope to encourage everyone to engage in cross-cultural considerations to make NLP more culturally inclusive.

Collaborative AI Using Semantic Communication and Challenges for the Next Generation AI

Yoshikazu NAKAJIMA
Professor of Department of Biomedical Informatics,
Institute of Science Tokyo, Japan

Abstract:
We proposed an artificial intelligence (AI) that connects automatic network information for collaboration and attempted to implement the self-organization of data processing networks for the purpose of medical data processing. Presently, most AI is purpose-specialized AI. On the other hand, generalization of AI is also being actively attempted, for example, by developing huge network models and devising deep learning methods for them. However, both theory and experiment have shown that excessive generalization of integrated networks reduces the accuracy of AI inference, and it shows a trade-off between high accuracy and generalization.

We adopted an approach to achieve both high accuracy and generality by preparing collaboration options with large number of highly-accurate purpose-specialized AIs, and self-organizing the data processing system while appropriately identifying and selecting them according to the purpose of the required data generation. We call this mechanism and implementation "Collaborative AI" that autonomously connects data and algorithms by autonomously-and-recursively identifying, matching, and collating semantics to self-organize an adequate data-processing network. Collaborative AI uses semantic communication to recursively match, select, and deploy semantic descriptions between inputs and outputs of goal-specific AI, which corresponds to "algorithms," to self-organize a data processing system that achieves its goals. In addition, we have developed Interface AI, Intelligent Database, and powerful Algorithm AI, including Vision AI, which recognizes human behavior through camera images, and Auditory AI, which recognizes human conversation through

microphone sounds, By connecting them with Collaborative AI, we implemented Artificial General Intelligence (AGI), which realizes generalization. In our experiments, we tested the self-organization and processing performance of the data processing system using medical data from seven large computers and 254 edge computers with a total of more than 12,000 tensor cores, which were used for surgical navigation and other applications. Compared to the combinatorial computation time of recursive algorithms in ordinary databases, which was O(N^2), and even in object-oriented databases, which was O(N log(N)), the computation time in our general-purpose AI was constant, independent of the increase in the number of possible combinations.

The ChatGLM's Road to AGI

TANG Jie
Professor
Tsinghua University

Abstract:
Large language models have substantially advanced the state of the art in various AI tasks, such as natural language understanding and text generation, and image processing, multimodal modeling.

In this talk, we will first introduce the development of AI in the past decades, in particular from the angle of China. We will also talk about the opportunities, challenges, and risks of AGI in the future. In the second part of the talk, we will use ChatGLM, an alternative but open sourced model to ChatGPT, as an example to explain our understandings and insights derived during the implementation of the model.

From Principle to Practice: The AI-45° Law for Balancing AI Safety and Capability

ZHOU Bowen

Abstract:
This talk introduces the "AI-45° Law" as a theoretical framework for balancing AI safety and capability. It presents a novel three-layer causal ladder theory for trustworthy AGI, comprising reflectable, intervenable, and approximate alignment layers, establishing a robust theoretical foundation for safe AGI development.

The presentation analyzes the synergistic development mechanism between AI capabilities and safety measures, proposing and validating the AI-45° Law. Through multiple empirical studies, including breakthroughs in safety alignment techniques and innovative approaches to multi-agent system safety evaluation, the effectiveness of this theoretical framework is demonstrated. The talk also introduces a new paradigm of AI evaluation based on causal reasoning and technical innovations in open-source evaluation platforms like OpenCompass. These theoretical and practical explorations provide new scientific insights for constructing a global AI safety framework.

Reliable AI System and Future

Kensaku MORI
Professor
Graduate School of Informatics, Nagoya University

Abstract:
In this talk, we will discuss reliable artificial intelligence (AI) for future collaboration between humans and AI systems. AI technology is now widely integrated into society, from GPT-based chat systems that enhance productivity to image recognition applications for person identification and autofocus in cameras. In the medical field, AI applications are advancing rapidly, including automated diagnoses of endoscopic images and detection of anomalies in X-ray and CT scans. In this context, developing reliable AI is crucial—not only in creating robust systems but also in defining what "reliable AI" means. AI reliability encompasses various aspects, including accuracy, performance, bias, data sources, and the risk of AI hallucinations. Based on our research, we will present a framework for defining AI reliability and showcase examples of reliable AI applications.

International Dialogue on the AI Governance

HARAYAMA (Yuko)
Professor Emeritus
Tohoku University

Abstract:
AI is becoming a common digital tool in our everyday lives and is expanding its reach into our society. Coupled with the advances in computing power and the ever-increasing ability to collect and create datasets, AI is challenging some human capabilities, while we are still far from ensuring the liability of the outputs it generates.

To accompany the technological development of AI and to ensure its responsible use of AI, the need for a governance framework or so-called “guardrails” has been advocated by both governments and the research community.

In my presentation, I will give a brief overview of the existing international fora in this endeavor and try to identify some initiatives for international cooperation.

AI Safety Issues for AI Sustainability

Young-Im CHO
Professor
Dept at Computer Engineering, Gachon University

Abstract:
This seminar will focus on examining the essential concerns and definitions that should shape our understanding of AI safety, a concept that currently lacks a standardized definition. Through a careful review of existing standards addressing trustworthiness, bias, risk management, and ethical and societal concerns, we will explore potential frameworks for AI safety that support sustainable AI development.

First, ISO/IEC 24028 on trustworthiness provides guidance on establishing essential elements such as transparency, explainability, and robustness in AI systems. Trustworthiness is a critical factor for AI systems, as it helps secure societal acceptance and enhances the long-term sustainability of AI technology.
Next, ISO/IEC 24027 focuses on addressing bias in AI algorithms and datasets, emphasizing the importance of fairness and impartiality across diverse social groups. Bias mitigation is key to creating AI systems that support inclusive and fair outcomes, contributing to sustainable AI that respects diverse demographic needs.

The third standard, ISO/IEC 23894, guides risk management in AI by providing a framework for assessing and mitigating potential risks associated with AI systems. Effective risk management is crucial in minimizing unintended consequences or malfunctions, thus ensuring the safe and reliable operation of AI.

Lastly, ISO/IEC 24368 deals with ethical and societal concerns, underlining values such as fairness, accountability, and transparency. This standard aims to align AI development with societal values and improve the social acceptance of

AI by addressing ethical and social impacts.

By reviewing these standards, this seminar will outline key concerns—such as trustworthiness, bias mitigation, risk management, and ethical considerations—that are essential for defining AI safety. Through this exploration, we aim to develop a foundational understanding of AI safety issues that will contribute to fostering AI systems that are safe, responsible, and sustainable for society.

Policy Measures and International Collaboration in AI

Kyunghee SONG
Head
Research Center for AI at Sungkyunkwan University

Abstract:
The rapid development of artificial intelligence technologies presents both opportunities and challenges for countries worldwide. In Northeast Asia, Korea, China, and Japan each possess unique strengths in AI research and innovation, yet face similar obstacles related to regulation, ethics, and safety. This presentation outlines policy measures and collaborative strategies aimed at enhancing AI governance through a multi-faceted approach that begins with academic cooperation and evolves toward government-level coordination.

Central to this strategy is the emphasis on academic collaboration, led by national engineering academies. Regular academic forums and joint research initiatives can serve as a platform for sharing insights on critical issues such as algorithm bias, data privacy, and AI ethics. These cooperative efforts not only foster mutual understanding but also lay the groundwork for harmonized standards that can be adopted at the policy level.

Joint research and development initiatives further strengthen these academic ties by leveraging each country's expertise to address shared challenges. Collaborative projects and academic-industry partnerships can facilitate the testing and application of new technologies, ensuring that AI systems are transparent, fair, and secure. Workshops and training programs focused on AI policy implications and crisis management strategies are proposed to build readiness for future policy implementation and crisis response.

To bridge the gap between academic insights and policy action, the proposal includes creating pathways for government participation. Initial observer roles

in academic forums can build trust and inform future policy alignments, while academic policy recommendations can provide a structured roadmap for government discussions. This phased approach aims to establish a multilateral AI governance framework, centered on coordinated policy and regulatory harmonization across the three nations.

The presentation concludes by emphasizing that while establishing large-scale AI governance frameworks may face practical challenges, starting with strengthened academic collaboration offers a viable foundation. Such academic-led initiatives can incrementally pave the way for deeper government cooperation, ultimately fostering a cohesive and effective AI governance ecosystem in Northeast Asia.

Challenges and Opportunities for Global AI Governance

ZENG Yi
Professor, Chinese Academy of Sciences

Abstract:
In this talk, I will firstly discuss major challenges for AI Governance both for the near term and for the long term. I firstly discuss from a governance perspective, what AI should be used and should not be used for. Then I introduce our analysis on unbalanced and missing efforts for development and use of AI in major sustainable development related areas. Then I focus on safety and ethics for foundational models in the near term, and then to catastrophic risks and how we should get prepared. I argue a synthetic socio-technical perspective is truly and urgently needed. I also analyze the current landscape of AI Safety and Governance institutions and point out what is needed to deal with the bottle neck and make a progress for good, and for all. Then I will discuss major visions for AI in current and future society from different cultures and how we should get prepared from both AI and human perspectives, for an inclusive and symbiotic society.