International Trends in AI Safety and Governance

On March 28, 2024, the University of Tokyo's Institute for Future Initiatives and the University of Tokyo's Tokyo College held a public event titled “International Trends in AI Safety and Governance.” The event was held at the SMBC Academia Hall in the International Academic Building at the Hongo Campus of the University of Tokyo and was attended by 178 people online and about 30 people at the venue.

While discussions on the safety of AI have been developing domestically and internationally with the spread of generative AI, there are various types of discussions on "safety" and their countermeasures. In addition, with the AI Safety Institute being established in the U.K., U.S., and Japan, it is important to understand the types of "safety" and countermeasures specific to Japan as a basis for future international collaboration. The event was attended by experts on AI governance from overseas, and discussions were held on international AI safety and governance trends.

Speakers and Panelists

Merve Hickok: President and Research Director at Center for AI & Digital Policy (CAIDP)

Cyrus Hodes: Lead, SAFE project at the Global Partnership on AI (GPAI)

Inma Martinez: Chair of the Multi-stakeholder Experts Group, Global Partnership on AI (GPAI)

Michael Sellitto: Head of Global Affairs at Anthropic

Yoichi Iida: Special Negotiator for Information and Communications International Strategy, International Strategy Bureau, Ministry of Internal Affairs and Communications

Hideaki Shiroyama: Professor, Institute for Future Initiatives, The University of Tokyo

Arisa Ema: Associate Professor, Tokyo College, University of Tokyo (Moderator)

(1) Opening remarks

Professor Hideaki Shiroyama of the University of Tokyo's Institute for Future Initiatives first gave opening remarks. Focusing on the governance of emerging technologies, Professor Shiroyama explained how the Institute's Technology Governance Research Unit has contributed to international discussions on AI safety through its research on Risk Chain Models and participation in the GPAI. In light of

recent rapid changes such as the explosive spread of AI and the G7 Hiroshima AI Process, he expressed his hope that this event will serve as a catalyst for discussion in the context of Japan, given the current need to organize issues related to the safety of AI and to develop a system to address them in a manner that is relevant to each site.

(2) Topics from presenters

First, as an introduction to the discussion points from the panelists, Ms. Inma Martinez of the GPAI mentioned the GPAI's emphasis on equity and inclusion of vulnerable peoples as "AI for all," and the leadership role Japan has played in the GPAI discussion, including these issues. In addition, she introduced that the most recent GPAI activities have been conducted with an emphasis on consensus building toward the realization of common values, a feature of the G7 Hiroshima AI process.

Ms. Martinez also explained that AI is not limited to automation, but will impact and transform all industrial sectors, and that while the "safety" of such AI can be interpreted in culturally diverse ways around the world, the "trustworthiness" of such AI is "technically functional," and that there is a consensus. It was then emphasized that while the GPAI seeks to build consensus, definitions on AI are no longer valid and should be in line with each country's culture and values and should not be monocultured.

Next, Mr. Cyrus Hodes, also from GPAI, stated that GPAI is working with multi-stakeholders to guarantee the safety of generated AI.

Mr. Hodes then noted that one of the risks of generative AI is that as AI systems become more sophisticated, which is bringing risks of misalignment, control and robustness of these systems and where tools addressing these raising issues will become increasingly important (such as audits, evaluations, cybersecurity red-teaming) and where an infrastructure for such alignment needs to be established, he expressed hope for collaboration with the AI Safety Institute. In addition, he mentioned that he expects Japan to cooperate in mapping the various set of tools developed by the global community and contribute to international coordination on AI safety.

Ms. Merve Hickok of CAIDP then spoke, first explaining that the Center is tasked with providing recommendations on AI policy to governments and international organizations, and training of future AI policy leaders. She then introduced the current state of AI policy in the U.S., which is consistent across Administrations, and the development of binding presidential executive orders for government agencies

and voluntary guidelines that can also be used in the private sector. She also explained that the bipartisan agreement on the need for AI regulation in the U.S. is a reflection of the failure to regulate harmful impact of social media. She noted the AI Safety Institute was established in the U.S., and that which ministry is in charge of this type of organization indicates what the nation is focusing on, she stated that in the U.S., unlike in UK, the definition of "safety" is broad and includes the economy and current risks of AI, and therefore, the Department of Commerce is in charge of this type of organization. In addition, she introduced recent initiatives such as the AI Safety Summit by the UK, upcoming AI Summit in France, and the Mini-Virtual Summit in South Korea.

Ms. Hickok emphasized the importance of "interoperability" to avoid governance fragmentation. However, she also warned about reducing the protections to a minimum number of common elements in the discussion of AI and human rights. She underlined the importance of international collaboration with multi-stakeholder participation, and advancing the elements of the Hiroshima AI process.

Finally, Mr. Michael Sellitto of Anthropic introduced the company's Responsible Scaling Policy, introduced that under the Responsible Scaling Policy, an AI Safety Levels (ASL), analogous to the biosafety level, is set and safety and security measures are taken according to the degree of risk. He also said that calls for a moratorium on AI development should not be based on abstract risks, but should be considered only when there is concrete evidence that safety or security measures may be insufficient.

Mr. Sellitto also praised the international code of conduct developed during the Hiroshima AI process as a highly effective framework, and expressed hope that the public and private sectors will work together to monitor commitments and thus increase confidence in the code.

(3) Panel Discussion

Following the introduction of the above issues, Mr. Iida, Special Negotiator for International Information and Communications Strategy, International Strategy Bureau, Ministry of Internal Affairs and Communications, and Prof. Shiroyama joined a panel discussion moderated by Associate Prof. Ema on the topic of “What is expected of Japan in AI governance.”

First, Mr. Iida expressed his appreciation for the substantial presentations, including the introduction of overseas case studies, as well as his compliments for the

ambitious efforts by each stakeholder to address AI safety. Mr. Iida also stressed the importance of ensuring commonality and interoperability in the diversity of AI policies, while pointing out that even among developed countries, there is still diversity, especially in approaches, as indicated by the comments of each speaker. He also noted Anthropic's voluntary efforts and willingness for international collaboration, which he appreciated and felt reinforced by such efforts.

Professor Shiroyama posed the question as a discussant, what is safety and why is it important? He then asked for further views on two points: what are the new risks posed by advanced and generative AI that differ from conventional AI, and what does the existence of bipartisan consensus and differences in competent ministries mean when comparing AI policies in different countries?

In response to the issues raised by Professor Shiroyama, Mr. Sellitto first responded that while there are a wide range of concerns and risks surrounding AI, "safety" in the context of Anthropic's focus is to ensure that AI can be used in a reliable and safe manner.

Ms. Martinez then noted that the 21st century is the first century in which safety has been brought to all industries but pointed out that "safety" is about preventing harm, not causing harm.

In response, Ms. Merve pointed out that while the objective function of AI is the starting point for trust and performance evaluation, it is not possible to envision all use cases for general-purpose AI. This makes it harder to manage risk and performance.

Mr. Hodes also noted that in the era of AGI, any task can be subject to improvement by AI, but values must be maintained by adjusting AI systems in such a society.

In response to these issues, Mr. Iida explained how the Hiroshima Process was launched to discuss the risks of generative AI but was later added to cover infrastructure systems and advanced AI as well. He also recognized that in international discussions, "safety" and "trust" have been discussed simultaneously, and that discussions on the definition of safety have been avoided, and that a detailed definition is needed in the course of taking concrete measures in the future.

Associate Professor Ema, the moderator of the session, also raised the point that discussions on safety should be framed not only in terms of the safety of AI itself, but also in terms of the safety realized by AI, such as its use in law enforcement agencies, and the trade-off relationship with other values.

In response, Mr. Iida noted that both Professor Shiroyama's and Associate Professor Ema's points of view are extremely important, but also expressed the view that the gap between political and administrative actors in terms of minimizing risk while advancing technology-based innovation is not so large. Mr. Iida also reiterated the importance of a multi-stakeholder approach in the AI policy-making process.

Mr. Hodes agreed with Mr. Iida, pointing to the composition of the U.S. and China as the two giants, and praised Japan's efforts, such as the establishment of the AI Safety Institute, and expressed hope that Japan would play a coordinating role.

Ms. Merve, while noting the differences in authority among ministries, emphasized the importance of a multi-stakeholder approach because of the need for diverse competencies, and praised Japan’s work to drive commonalities across actors.

Ms. Martinez noted that even in Europe, the development of regulations pertaining to the Internet has been slow but said that regulations on AI have been developed under a global consensus based on principles, values, and commonalities, taking into account Japan's recommendations.

Mr. Sellitto noted that in the early stages of technology development, there can be concern that regulations will hinder innovation, but people will gradually learn what to regulate, and that Anthropic's ASL was also a practice of first developing and implementing commitments and then publishing the lessons learned from them, and he hopes that this will lead to the development of best practices that can inform regulations in the future.

(4) Questions from an audience

In response to a question from an online participant about what is needed to ensure the safety and reliability of AI, given that Japan has been the target of cyber-attacks in recent years, Mr. Sellitto explained that while there are currently no clear guidelines for AI cyber security, he explained that cybersecurity standards are being formed. Ms. Martinez also expressed the view that there have been many cyberattacks targeting AI, we can learn from them to increase resilience.

(5) Summary and closing remarks

In concluding the event, Professor Shiroyama summarized the discussions and pointed out the need to organize a common vocabulary and know-how for "safety," although it seems better not to dare to establish a detailed definition. He also suggested that the dichotomy of hard law/soft law for regulation of AI is too

simplistic, and that the learning process needs to start with abstract principles and shared experiences.

In addition to thanking the participants, Associate Professor Ema mentioned the need to adhere to an agile process for AI security and safety, and ultimately AI governance, in the face of rapid technological innovation.

Finally, Prof. Takeo Hoshi, Deputy Director of Tokyo College at the University of Tokyo, gave closing remarks. Prof. Hoshi pointed out the importance of today's discussion, and expressed the pleasure for Tokyo College to host this event together with the Institute for Future Initiatives. Drawing on debates on regulatory attempts to prevent financial crises, which is one of his areas of expertise, he stated that financial crises have been happening despite the various efforts to build sound and safe financial systems. here seem to be no regulatory mechanisms that make the financial systems completely safe. The lesson is that, in addition to trying to prevent crises, we need to be ready to respond. Prof. Hoshi concluded the event by noting the need to prepare for AI crises while promoting human-centered AI development, and expressed his hope that today's discussion would serve as a starting point for future discussions.