The Path to Trustworthy AI: G7 Outcomes and Implications for Global AI Governance

Photo: JONATHAN ERNST/POOL/AFP via Getty Images

Commentary by **Hiroki Habuka**

Published June 6, 2023

Introduction

The G7 Summit, held from May 19 to 21, 2023, left a strong impression of unity among global leaders. Emerging technology such as artificial intelligence (AI), along with national security and energy, was highlighted as a key area requiring a strong alliance among G7 countries toward fundamental values such as democracy and human rights. The G7 Digital and Tech Ministers' Meeting, which took place a little earlier, also discussed responsible AI and global AI governance as one of the central topics.

With the recent launch of large language AI models such as GPT-4, society anticipates a rapid integration of AI technologies, making it crucial to discuss the responsible use of AI and its governance. However, even the like-minded G7 countries approach AI governance differently, ranging from strict, comprehensive laws to sector-specific guidelines. Because of this, the consensus among G7 leaders on trustworthy AI will have a big impact on the way AI systems should be developed and operated around the world. This commentary takes a close look at two important outcome documents from the 2023 G7 summit that broach the subject of AI governance: the G7 Summit Communiqué and the Declaration of the G7 Digital and Tech Ministers' Meeting–collectively referred to as the G7 statements–to predict potential cooperation on AI governance among the G7 countries.

Shared Understanding of Values, Principles, and Risks

1. Establishing Common Values

   The G7 is a forum of nations that share fundamental values, such as freedom, democracy, and human rights. Therefore, the G7 statements emphasize several times that AI development and implementation should be aligned with these values. The G7 statements also recognized the importance of basic principles such as fairness, accountability, transparency, and safety, mirroring those listed in the Organization of Economic Cooperation and Development (OECD) AI Principles. One highlight from the communiqué would be the focus on implementing procedures to promote the key principles. The communiqué recognizes "the importance of procedures that advance transparency, openness, fair processes . . . to promote responsible AI." The OECD AI Principles also include some procedural recommendations for AI actors, such as transparency and systematic risk management, but the openness or fair process in general were not explicitly addressed. Although the definitions of openness or fair process are open to interpretation, as stated several times in the communiqué, inclusive and multistakeholder dialogue would be the key component not only in policymaking but also in the development and implementation of AI systems.

2. Addressing Key Risks

   The communiqué outlines potential risks with AI, including online harassment, hate and abuse, and threats to children's safety and privacy. In the realm of generative AI, it points to the danger of foreign information manipulation, which involves spreading disinformation. The declaration extends this list, warning of AI's misuse and abuse in a way to undermine democratic values, suppress freedom of expression, and threaten the enjoyment of human rights. The risks posed by AI to freedom of expression entail, among others, excessive content blocking and restriction and opaque dissemination of information. In short, G7 statements express stronger concern about the potential for AI to produce and disseminate harmful content that would jeopardize foundational values such as human rights and freedom, rather than about AI's inherent risks, such as lack of predictability or explainability.

In addition, G7 statements also address the importance of international collaboration in protecting intellectual property (IP) rights, including copyrights, and promoting transparency in generative AI. The "Hiroshima AI Process" will be launched for discussions on generative AI by the end of this year in an inclusive manner and in cooperation with the OECD and Global Partnership in AI (GPAI).

The Road Ahead: Achieving Interoperability in AI Governance Frameworks

To maximize the shared values while mitigating the aforementioned risks, the G7 statements highlighted the necessity of interoperability among AI governance frameworks to foster trustworthy AI. This list outlines what exactly such a framework entails:

1. Risk-Based and Forward-Looking Approaches under Different Policy Frameworks

   First, it is important to note that the G7 statements do not attempt to harmonize the approaches and policy instruments of the G7 members, declaring that "the common vision and goal of trustworthy AI may vary across G7 members." Furthermore, the declaration emphasizes that policies and regulations should take into account not only technical and institutional characteristics, but also social and cultural impacts, including geographic, sectoral, and ethical aspects. In fact, there are policy differences among G7 members, from the European Union (France, Germany, and Italy) and Canada, which promote comprehensive and binding regulations for AI, to Japan, the United Kingdom, and the United States, which promote sector-specific guidance-based policies.

Under this fragmentation in AI governance approach, the G7 members agreed that policies and regulations should be risk-based and forward-looking. However, policymakers should be cautious about using the term "risk-based." The EU AI Act defines risk-based as the classification of AI systems into four categories: unacceptable, high-risk, limited, or minimal/low, depending on the purposes and actors using the system (e.g., social scoring by governments is classified as "unacceptable," while critical infrastructures that could put the life and health of citizens at risk are classified as "high risk"). In contrast, the U.S. NIST AI Risk Management Framework assumes that risk is assessed for each AI system based on the magnitude of risk and probability of occurrence. In short, the term risk-based varies from country to country, and its common meaning may not differ significantly from the general principle of regulation: necessary and proportional.

2. Partnership with International Organizations and Multi-stakeholder Initiatives

The G7 statements emphasize the need to support the development of tools for trustworthy AI several times. Tools here include a wide range of regulatory and nonregulatory frameworks, technical standards and assurance techniques, risk assessment and management frameworks, auditing, and potential certification schemes.

The first expected drivers of such tools are international organizations such as the OECD and United Nations Educational, Scientific and Cultural Organization (UNESCO), as well as multi-stakeholder initiatives such as the GPAI. In particular, the OECD is the organization mentioned most often in the G7 statements, not only as the issuer of the OECD AI Principles, but also as the research hub such as mapping the commonalities and differences between trustworthy AI frameworks. UNESCO published a document, Recommendation on the Ethics of Artificial Intelligence, in 2021, and various projects are underway to put it into practice. GPAI is a multi-stakeholder, academia-led initiative that aims to bridge the gap between theory and practice on AI, including the research on use of privacy enhancing technologies, the use of AI to solve environmental problems, and many other projects.

The frameworks and tools created by these international organizations and initiatives will serve as benchmarks that countries can refer to, for example, in considering regulatory frameworks or allocating civil liabilities.

3. Support for International Technical Standards

The other approach stressed in the G7 Statements to ensure interoperable AI governance framework is the support for international technical standards by standards development organizations (SDOs).

Technical standards function as a baseline to gauge a product's features and performance. For example, ISO/IEC JTC 1/SC 42, an international standard for AI, includes “assessment of machine learning classification performance” and “big data reference architecture.” These technical standards enable a common platform for risk assessments and audits, allowing countries with varying regulations to mutually assess and evaluate AI systems or services. It is worth pointing out, however, that technical standards are not a panacea. To create a truly interoperable framework for AI governance, collaboration that goes beyond just technical aspects and includes socio-technical and normative elements is needed as well. For instance, in the NIST AI Risk Management Framework's Initial Draft, aspects such as explainability and bias are labeled as socio-technical elements, and fairness, accountability, and transparency are referred to as “guiding principles.” These elements, distinct from purely technical components, should be implemented through continuous interaction with stakeholders and coordination of benefits and risks throughout the AI product's lifecycle.

Another point to consider is that the G7 statements do not explicitly address the link between regulations and standards. Even if international standards are established, unless they are aligned with each country's regulatory content and civil liability systems, businesses could still face challenges due to regulatory inconsistencies. This suggests a need for ongoing discussions among nations, not just to support the development of standards, but also to understand the interplay between regulations and standards.

Further Collaboration in DFFT and Emerging Technologies

This commentary analyses the AI-focused sections ("Digital" in the communiqué and "Responsible AI and Global AI Governance" in the declaration), but they are not the only points concerning AI governance. Other sections, specifically those on Data Free Flow with Trust (DFFT) and Emerging and Disruptive Technologies in Innovative Society and Economy, offer additional valuable insights into AI governance. This list will delve into these sections and distill their key implications for AI governance.

1. Institutional Arrangement for Partnership

   The performance of AI systems is greatly determined by data. To allow people worldwide to reap the benefits of AI systems, the free flow of high-quality data is essential. However, worldwide data policies significantly diverge, covering domains such as privacy, government access, security, harmful content, and IP rights. Initially brought forward by former prime minister Shinzo Abe at Davos in 2019, the DFFT initiative seeks to enhance the cross-border flow of data that is beneficial for business and social problem-solving. This goal is pursued while simultaneously ensuring trust in privacy, security, IP rights, and other associated areas. Commitments to the DFFT have been reiterated at the Japan 2019 G20, the UK 2021 G7, and the Germany 2022 G7. However, no specific collaborative framework has been established so far to achieve this goal.

   This time, The G7 statements concurred on the establishment of an Institutional Arrangement for Partnership (IAP) to operationalize DFFT. This is a principles-based, solutions-oriented, evidence-based, multi-stakeholder and cross-sectoral cooperation initiative. Being solution-oriented suggests that this initiative focuses not on specific concepts such as privacy or fairness but on practical solutions to actual problems. More specifically, the project is expected to deal with improving accessibility to regulatory information, cooperation in privacy-enhancing technologies, model contractual clauses, digital credentials, and identities. The OECD is anticipated to play a leading role in advancing the IAP.

These endeavors have a direct link to trustworthy AI. For instance, having trust in the dataset used to train an AI system is crucial to ensuring its transparency and accountability. Moreover, the question of how to build trust in data generated by AI systems is an inevitable challenge in realizing the DFFT. Discussions on how to make such trustworthy AI ecosystems will be promoted under the umbrella of the IAP in partnership with the public and private sectors.

2. Implementing Agile, Distributed, and Multi-stakeholder Governance through Policy Incentives

   Another important section related to trustworthy AI is the Emerging and Disruptive Technologies in Innovating Society and Economy. In this section, digital and technology ministers emphasized the need for agile, more distributed, and multi-stakeholder governance and legal frameworks for operationalizing the principles of the rule of law, due process, democracy, and respect for human rights. These deliberations will be based on input from the Taskforce on Governance for a Digitalized Society (TGDS).

   The TGDS is an expert group of leading academics and practitioners from G7 member countries. It proposed the "Governance Principles for a Society Based on Cyber-Physical Systems" framework, which emphasizes the need for an agile, distributed and multi-stakeholder process (so-called Agile Governance) for Cyber-Physical Systems with AI at its core. The report further suggested a policy package to promote Agile Governance, that includes (1) proactive governance by organizations and individuals that deploy or operate aspects of CPS, (2) expert involvement and utilization of digital tools, (3) agile regulatory governance, (4) reliable certification mechanisms, (5) effective enforcement systems and appropriately tailored liability systems, and (6) legal remedial measures. Such an integrated policy framework will provide an important perspective for the development of an interoperable governance framework for responsible AI.

Conclusion

The G7 statements have highlighted the shared commitment to developing and implementing trustworthy AI that upholds values such as human rights and democracy. The statements also acknowledge risks including online abuse, threats to privacy, misuse of AI, and IP concerns. The G7 leaders have agreed to support the establishment of an interoperable governance framework to counteract potential gaps and fragmentation in global technology governance. They have also committed to assisting the efforts by organizations such as OECD and GPAI, and to collaborate in the development of international technical standards under SDOs.

Yet it is not entirely clear what kind of cooperation might extend beyond the existing national policies and partnerships with international initiatives. The additional sections of the G7 statements that are closely related to AI governance give some insights.

In the DFFT section, the creation of the new IAP has been agreed on. This arrangement would be rooted in multi-stakeholder problem-solving, guided by shared principles. Further, in the section on "Emerging and Disruptive Technologies in Innovating Society and Economy," the necessity of an agile, dispersed, and multi-stakeholder governance approach for AI-driven cyber-physical systems is recognized, and a policy package has been suggested to actualize this.

As the word "stakeholder" frequently appears in the G7 statements, the future of AI governance will likely not rely solely on top-down, government-led rule-setting. Instead, it suggests an approach where multi-stakeholder initiatives are discussed on a case-by-case basis and updated in an agile manner. Regulations and technical standards serve as critical tools to facilitate these efforts in building trustworthy AI. However, it is crucial to note that the key players in AI governance, who develop and use these tools, will be not only the government entities, but rather the private stakeholders including AI developers, users, and civil society organizations, and so on.

Principles for trustworthy AI such as fairness, accountability, and transparency should be realized through ongoing multi-stakeholder dialogues, as these principles are subject to the ever-changing geographical, sectoral, and ethical contexts. The G7 statements play a crucial role in setting the direction for such multi-stakeholder, distributed, and agile governance for trustworthy AI.

Hiroki Habuka is a non-resident fellow with the Wadhwani Center for AI and Advanced Technologies at the Center for Strategic and International Studies in Washington, D.C.

Commentary is produced by the Center for Strategic and International Studies (CSIS), a private, tax-exempt institution focusing on international public policy issues. Its research is nonpartisan and nonproprietary. CSIS does not take specific policy positions. Accordingly, all views, positions, and conclusions expressed in this publication should be understood to be solely those of the author(s).

Tags

Technology Global Economic Governance and Artificial Intelligence

Center for Strategic and International Studies
1616 Rhode Island Avenue, NW
Washington, DC 20036

Tel: 202.887.0200
Fax: 202.775.3199

MEDIA INQUIRIES

H. Andrew Schwartz

Chief Communications Officer

☎ 202.775.3242

✉ aschwartz@csis.org

Samuel Cestari

Media Relations Coordinator, External Relations

☎ 202.775.7317

✉ scestari@csis.org

See Media Page for more interview, contact, and citation details.