Japan's Efforts toward Global Governance of AI

~ Hiroshima AI Process~

Jun. 2024

Yoichi IIDA

Assistant Vice Minister for International Affairs, MIC

Chair, Committee on Digital Economy Policy, OECD

Chair, Hiroshima AI Process WG at G7

Launch of the Hiroshima AI Process (HAIP) at G7

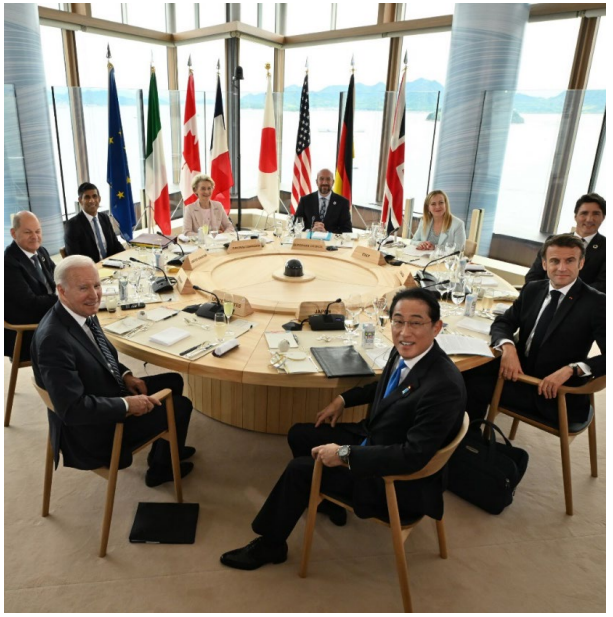
How it started

November 30, 2022
Chat GPT 3.5 prototype release by Open AI

March 15, 2023
Chat GPT 4 release

April 29-30, 2023
G7 Digital and Technology Ministerial Meeting (Takasaki, Gunma)

May 19-21, 2023
G7 Hiroshima Summit



G7 Hiroshima Leaders' Communiqué Excerpts

We **recognize the need to immediately take stock of the opportunities and challenges of generative AI**, which is increasingly prominent across countries and sectors, and encourage international organizations such as the OECD to consider analysis on the impact of policy developments and Global Partnership on AI (GPAI) to conduct practical projects. In this respect, **we task relevant ministers to establish the Hiroshima AI process, through a G7 working group**, in an inclusive manner and in cooperation with the OECD and GPAI, **for discussions on generative AI** by the end of this year.

History of Discussions on AI principles

2016	2017	2018	2019	2020	2021	2022
G7 ICT Ministers' Meeting (Japan)	G7 ICT and Industry Ministerial Meeting (Italy)	G7 Innovation Ministers' Meeting and Summit (Canada)	G7 Digital Ministerial Meeting (France)	G7 Science Technology Ministerial Meeting (US)	G7 Digital and Technology Ministers' Meeting (UK)	G7 Digital Ministers' Meeting (Germany)
Proposal of Discussion on AI R&D Guideline	Sharing the vision of "Human-centric AI"	Common vision for the future of AI	G20 Ministerial Meeting on Trade & Digital Economy (Japan) G20 Osaka Summit	Launch the Global Partnership on AI (GPAI)		GPAI Summit 2022(Tokyo)

G20 AI Principles *(June.2019)*

GPAI

Project based

OECD

Recommendation on Artificial Intelligence *(May.2019)*

OECD AIGO

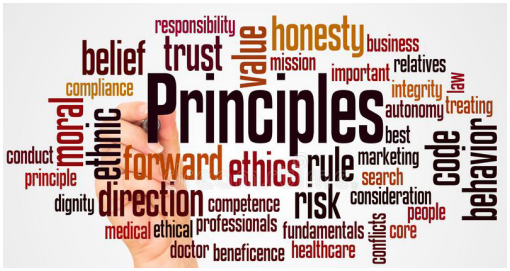
Risk assessment

Non-binding regulatory framework
Promoting a human-centric approach to trustworthy AI,
Shaping an enabling policy environment

EU
2022~2023 : regulation in a transitional period
2024 : regulation could become applicable

Enacting AI ACT

The OECD AI Principles



5 values-based principles for trustworthy, human-centric AI

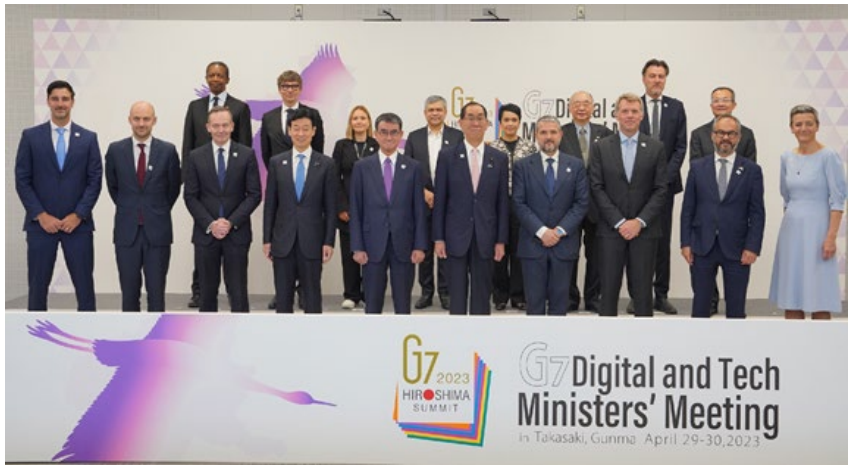
- **Benefit People & Planet**
- **Human rights, values & fairness**
- **Transparent & explainable**
- **Robust, secure & safe**
- **Accountable**

5 recommendations for national policies, for AI ecosystems to benefit societies

- AI research & development
- Data, compute, technologies
- Policy & regulatory environment
- Jobs & skills, labour transitions
- International cooperation & measurement

AI discussion in G7 Digital & Tech Track in 2023

G7 Digital and Tech Ministers Meeting Declaration (30 Apr 2023) : Excerpt



- Stress **the importance of international discussions on AI governance and interoperability between AI governance frameworks**, while we recognise that like-minded approaches and policy instruments to achieve the common vision and goal of trustworthy AI may vary across G7 members.

- Plan to **convene future G7 discussions on generative AI** which could include topics such as governance, how to safeguard intellectual property rights including copyright, promote transparency, address disinformation, including foreign information manipulation, and how to responsibly utilise these technologies.

HAIP – Overview of Progresses and Way Forward

"Hiroshima AI Process" was launched in May 2023.

- ☐ On October 30, the "G7 Leaders' Statement on the Hiroshima AI Process" delivered **Hiroshima Process International Guiding Principles (GPs)** and **International Code of Conduct (CoC) for Organizations Developing Advanced AI System**.
- ☐ On December 6, 2023 **G7 Leaders welcomed the "Hiroshima AI Process Comprehensive Policy Framework"** as outcomes from Hiroshima AI Process.
- ☐ In 2024, the Italian Presidency continues work to advance Hiroshima AI process, particularly in **developing monitoring mechanism** for CoC, and in May, **Hiroshima AI Process Friends Group**, was launched with 49 countries and regions.

Japan's presidency in 2023 | **Italy's presidency in 2024**

May

G7 Hiroshima Summit

Launch of Hiroshima AI Process

Sep

G7 Ministerial Meeting

Agreement on Ministers' Statement

Oct

Multi-stakeholder High-Level Meeting

Share Discussion of Hiroshima AI Process

Dec

G7 Leaders' Statement

Approval of the achievement of the Ministerial Meeting

G7 Ministerial Meeting

HAIP Comprehensive Policy Framework including the International Guiding Principles for all AI actors, International Code of Conduct for AI developers and project-based cooperation, and a work plan.

Mar

Implementation of International Code of Conduct

G7 Industry, Technology and Digital Ministerial Meeting

May

OECD Ministerial Council Meeting

Launch of the Hiroshima AI Process Friends Group

Expansion of supporting countries

Hiroshima AI Process G7 Digital and Technology Ministerial Statement (Dec 1, 2023)

- Hiroshima AI Process Comprehensive Policy Framework -

The G7 endorsed the Hiroshima AI Process Comprehensive Policy Framework, comprising of the following 4 elements, as the first successful international framework addressing the impact of advanced AI systems including generative AI on our societies and economies .

1. OECD's Report towards a G7 Common Understanding on Generative AI

- **Key Areas of Concern** ; lack of transparency, disinformation, intellectual property rights, privacy and protection of personal data, fairness, security and safety, amongst others.
- **Opportunities** ; productivity gains, innovation and entrepreneurship, healthcare, and the climate crisis

2. Hiroshima Process International GPs for All AI Actors

- 11 items agreed as Guiding Principles for Organizations Developing Advanced AI Systems, but should be also applied to **all AI actors** as appropriate.
- 12th item added to encourage AI actors to **improve digital literacy, training and awareness** and to **cooperate and share information** to address risks and vulnerabilities

3. Hiroshima Process International CoC for Organizations Developing Advanced AI System

- Code of Conduct presents **specific actions and measures** for implementation of individual principle items.
- The G7 **reach out to organizations** to encourage implementation of Code of Conduct.

4. Project-Based Cooperation

- The G7 collaborate in exploring **technological solutions** against challenges by advanced AI systems, such as the **spread of disinformation**.
- **Global Challenge project** coordinated by the OECD, IEEE
- Other possible projects may be supported by **GPAI Tokyo Center** and more

* GPs: Guiding Principles, CoC: Code of Conduct

Outline of the Hiroshima Process International Guiding Principles for All AI Actors

- The G7 emphasizes **the responsibilities of all AI actors in promoting**, as relevant and appropriate, **safe, secure and trustworthy AI**.
- The G7 **encourages all AI actors to read and understand** the "Hiroshima Process International Guiding Principles for Organizations Developing Advanced AI Systems (October 30, 2023)" **with due consideration to their capacity and their role within the lifecycle**.

1. Take appropriate measures throughout the development of advanced AI systems, including **prior to and throughout their deployment and placement on the market, to identify, evaluate, and mitigate risks across the AI lifecycle**

2. Identify and mitigate vulnerabilities, and, where appropriate, incidents and patterns of misuse, **after deployment including placement on the market**

Outline of the Hiroshima Process International Guiding Principles for All AI Actors

3. **Publicly report advanced AI systems' capabilities, limitations and domains of appropriate and inappropriate use**, to support ensuring sufficient transparency, thereby contributing to increase accountability

4. **Work towards responsible information sharing and reporting of incidents** among organizations developing advanced AI systems including with industry, governments, civil society, and academia

5. **Develop, implement and disclose AI governance and risk management policies**, grounded in a risk-based approach – including privacy policies, and mitigation measures, in particular for organizations developing advanced AI systems

Outline of the Hiroshima Process International Guiding Principles for All AI Actors

6. **Invest in and implement robust security controls, including physical security, cybersecurity and insider threat safeguards** across the AI lifecycle

7. **Develop and deploy reliable content authentication and provenance mechanisms, where technically feasible, such as watermarking or other techniques** to enable users to identify AI-generated content

8. Prioritize research to mitigate societal, safety and security risks and **prioritize investment in effective mitigation measures.**

Outline of the Hiroshima Process International Guiding Principles for All AI Actors

9. **Prioritize the development of advanced AI systems to address the world's greatest challenges**, notably but not limited to the climate crisis, global health and education

10. **the development of and, where appropriate, adoption of international technical standards**

11. **Implement appropriate data input measures and protections for personal data and intellectual property**

12. **Promote and contribute to trustworthy and responsible use of advanced AI systems**

improve their own and, where appropriate, others' **digital literacy, training and awareness**, including on issues such as how advanced AI systems may **exacerbate certain risks** (e.g. with regard to the **spread of disinformation**) and/or **create new ones**.
All relevant AI actors are encouraged to cooperate and **share information, as appropriate, to identify and address emerging risks and vulnerabilities** of advanced AI systems.

Outline of HAIP Code of Conduct for Organizations Developing Advanced AI Systems

Code of Conduct is formulated based on No1 to No11 of Guiding Principles, and presents specific actions and measures for implementation of individual principle items.

- Take appropriate measures prior to and throughout their deployment and placement on the market
  - ***Examples from CoC: Identification and mitigation of risks through internal and independent external testing, including "red teaming" prior to placement on the market***
- Mitigate vulnerabilities, incidents and patterns of misuse, after deployment including placement on the market.
  - ***Examples from CoC: Facilitating third-party and user discovery and reporting of issues and vulnerabilities after deployment such as through bounty systems or contests***
- Publicly report advanced AI systems' capabilities, limitations and domains of appropriate and inappropriate use
  - ***Examples from CoC: Publishing transparency reports which include, for example, details of the evaluations conducted for potential safety, security, and societal risks, as well as risks to human rights***
- Work towards responsible information sharing and reporting of incidents
  - ***Examples from CoC: Developing, advancing, and adopting, where appropriate, shared standards, tools, mechanisms, and best practices for ensuring the safety, security, and trustworthiness of advanced AI systems***
- Develop, implement and disclose AI governance and risk management policies, and mitigation measures.
  - ***Examples from CoC: Disclosing where appropriate privacy policies, including for personal data, user prompts and advanced AI system outputs***

Outline of HAIP Code of Conduct for Organizations Developing Advanced AI Systems

Ĭ Ḱ Invest in and implement robust security controls, including physical/cybersecurity and insider threat safeguards.

ṕ ***Examples from CoC Ṗ Securing model weights and, algorithms, servers, and datasets, such as through operational security measures for information security and appropriate cyber/physical access controls***

ĭ Ḱ Develop and deploy reliable content authentication and provenance mechanisms, such as watermarking or others

ṕ ***Examples from CoC Ṗ Developing tools or APIs to allow users to determine if particular content was created with their advanced AI system, such as via watermarks***

Į Ḱ Prioritize research to mitigate societal, safety and security risks and investment in effective mitigation measures.

ṕ ***Examples from CoC Ṗ Collaborating on and investing in research on upholding democratic values and respecting human rights***

į Ḱ Prioritize the development of advanced AI systems to address the world's greatest challenges

ṕ ***Examples from CoC Ṗ Supporting progress on the United Nations Sustainable Development Goals, and to encourage AI development for global benefit***

10 Ḱ Advance the development of and, where appropriate, adoption of international technical standards

ṕ ***Examples from CoC Ṗ Contributing to the development and, where appropriate, use of international technical standards and best practices, including for watermarking***

11 Ḱ Implement appropriate data input measures and protections for personal data and intellectual property

ṕ ***Examples from CoC Ṗ Implementing appropriate safeguards, to respect rights related to privacy and intellectual property, including copyright-protected content***

G7 Industry, Technology and Digital Ministerial Declaration (Extract) (Mar 14-15, 2024)

- Advancing the Outcomes of the Hiroshima Artificial Intelligence Process-

We **remain committed to advancing the Hiroshima AI Process outcomes**, including through **expanding support and awareness among key partners and organisations**, as well as increasing their involvement, as appropriate.

- Based on the Framework we look forward to continuing our work, including the **development of mechanisms to monitor the application of the Code of Conduct by organisations** that will commit to these outcomes on a voluntary basis, with the support of the OECD and informed by other stakeholders, organisations, and initiatives as relevant, such as UNESCO and the Global Partnership on AI (GPAI),

- We welcome the awareness raising event that took place on 22 January 2024 and **look forward to other opportunities for engagement with key partner countries**, including from developing countries and emerging economies, and organisations.

HAIP Friends Group

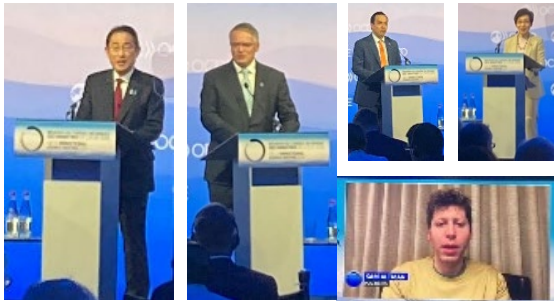
HAIP outcomes were widely supported by **49 countries/region**, and OECD adopted the **updated OECD AI Principles** at the Ministerial Council Meeting (MCM)

Side Event on Generative AI

As 2024 MCM Chair country, Japan hosted **a side event focused on HAIP**.

Prime Minister Kishida's message ;

1) Japan **celebrates the update of the OECD AI Principles**.
2) Japan **welcomes the additional countries who support the outcomes of the Hiroshima AI Process to promote safe, secure and trustworthy AI as part of the Hiroshima AI Process Friends Group. Japan will work with the 49 countries and a region in the Friends Group**.
3) Japan is launching **GPAI Tokyo Center** to provide support to projects to develop technological solutions with experts.



Also joined by ; Mr. Cormann, Secretary-General of the OECD,
Dr. Encinas, Undersecretary of Foreign Trade of Mexico,
Ms. Fu, Minister for Sustainability and Environment of Singapore
Mr. Altman, the CEO of OpenAI (online)

Session 6.1: Artificial Intelligence

OECD Member countries and guests, including private sector discussed **update of OECD AI principles.**

Many countries welcomed **productive synergy between HAIP and OECD Principles**.

Update of OECD AI principles was adopted. OECD AI Principles will continue to provide the robust foundation for international AI policy making.

Member countries of the Hiroshima AI Process Friends Group (As of June 30)

53 countries and regions have joined the Hiroshima AI Process Friends Group, a voluntary framework of countries supporting the sprit of the Hiroshima AI Process, to achieve safe, secure, and trustworthy AI.

(G7)

Canada	France	Germany	Italy	Japan
United Kingdom	United States	European Union		

(EU member countries)

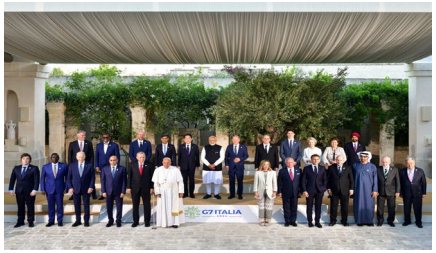
Austria	Belgium	Bulgaria	Croatia	Cyprus
Czech Republic	Denmark	Estonia	Finland	Greece
Hungary	Ireland	Latvia	Lithuania	Luxembourg
Malta	Netherlands	Poland	Portugal	Romania
Slovakia	Slovenia	Spain	Sweden	

(Others)

Argentine	Australia	Brunei	Chile	Colombia
Costa Rica	Iceland	India	Israel	Kenya
Republic of Korea	Laos	Mexico	New Zealand	Nigeria
Norway	Serbia	Singapore	Thailand	Türkiye
United Arab Emirates				

G7 Summit Leaders' Communique (Extract)

(June 13-15, 2024@ Apulia, Italy)



- Recognizing the importance of advancing the Hiroshima AI Process outcomes, we welcome **support from the countries and organizations beyond the G7, as demonstrated by its Friends Group.**

  We will step up our efforts to **enhance interoperability amongst our AI governance approaches** to promote greater certainty, transparency and accountability while recognizing that approaches and policy instruments may vary across G7 members.

- We are also committed to deepening coordination between our respective institutes and offices focused on AI, to **work towards shared understanding of risk management and advance international standards** for AI development and deployment.

- We welcome our Industry, Tech, and Digital Ministers' efforts to advance the Hiroshima AI Process outcomes released last year, including **the development of a reporting framework for monitoring the International Code of Conduct for Organizations Developing Advanced AI Systems.** We look forward to **the pilot of the reporting framework**, developed in cooperation with the OECD,

- We will work towards **developing a brand that can be used to identify organizations** that are voluntarily participating in and implementing the Code's forthcoming reporting framework.

Thank You !

If you have questions or comments;

y.iida@soumu.go.jp