【Provisional translation】

※当翻訳は仮訳であり、正確には原文を参照してください。
※Please refer to the original text for accuracy

Guideline for Ensuring the Appropriateness of Research & Development and Utilization of Artificial Intelligence-Related Technologies

(Draft)

【Provisional translation】

Table of Contents

【Provisional translation】

1. Basic approach to ensuring appropriateness in Japan

(1) Positioning of this guideline

This guideline, based on Article 13 of the Act on Promotion of Research & Development(R&D), and Utilization of Artificial Intelligence-Related Technologies (AI Act, Act No. 53 of 2025), is formulated in accordance with the spirit of international norms toward realization of trustworthy AI, aiming to encourage voluntary and proactive efforts by all stakeholders[1]—including businesses, citizens, etc.—for the appropriate implementation for R&D and utilization of AI,

The structure of this guideline is as follows: Section 1 presents the main elements and basic policies necessary for ensuring appropriateness in AI R&D and utilization for all stakeholders. Sections 2 and onward describe specific matters that each stakeholder should especially address, based on Section 1. All stakeholders are required to recognize and understand the main elements necessary for ensuring appropriateness. Furthermore, regarding the matters to be addressed, stakeholders are required to respond appropriately at a suitable level, taking into account their scale, position, and the risks posed by AI, as well as the technologies and knowledge available at the time.

Japan will develop a framework centered on this guideline as an international model for development, utilization and spread of trustworthy AI as well as promote international cooperation on construction of AI governance, continuously leading global discussions, based on real achievement[2] of leading the "Hiroshima AI Process" which is a framework to make an international rule regarding AI.

(2) Concept of ensuring appropriateness in this guideline

AI contributes to economic growth and the advancement of citizens' lives and it is important to promote its social implementation and innovation. However, AI also poses various risks: technical risks

---

[1] The term refers to the national government, local governments, research & development institutes, utilization business operators, and citizens, whose responsibilities are stipulated in Articles 4 through 8 of the AI Act.

[2] At the G7 Hiroshima Summit in May 2023, the "Hiroshima AI Process" was launched. As an outcome under Japan's G7 Presidency, the "Hiroshima AI Process Comprehensive Policy Framework" concerning the development and use of advanced AI systems were compiled. The Hiroshima AI Process Friends Group, a voluntary framework of countries and regions that support the spirit of the Hiroshima AI Process, was established in May 2024, with 60 countries and regions participating as of December 2025. Also, in February 2025, the "Reporting Framework" commenced official operations, and as of December 2025, 24 companies have submitted responses.

such as misjudgment and hallucination[3], social risks such as the generation and spread of disinformation or misinformation, promotion of bias or discrimination, use on crime, excessive dependence, infringement of privacy or property rights, increased environmental burden, employment or economic instability, and security risks such as cyberattacks. These risks may change with technological progress of AI, and unknown risks may emerge, and social tolerance levels for these risks may also change.

Therefore, in ensuring the appropriateness, this guideline does not define its single or absolute standard. Instead, under the consideration that each stakeholder is expected to voluntarily address according to the characteristics, a way of use and purposes of AI they research, develop and utilize, and their position and social roles, based on "The principles of Human-Centered AI Society" (decided by the Integrated Innovation Strategy Promotion Council, March 29, 2019), this guideline describes the following main elements to consider.

Main elements to consider

- Human-centricity
  - ✧ Respecting human dignity and fundamental rights and complying with laws
  - ✧ Respecting diversity and inclusion so that everyone can benefit from AI and aiming for inclusive growth by various mankind pursuing happiness
  - ✧ Making a final judgement by human on their own regarding scope and conditions for utilizing AI
- Fairness
  - ✧ Preventing and avoiding unjust bias or discrimination in society resulting from AI utilization[4]
- Safety
  - ✧ Ensuring that AI utilization does not harm life, body, property, etc[5]

---

[3] Generative AI refers to the phenomenon where it outputs information that differs from the facts in a plausible manner.

[4] This includes ensuring that fairness is not undermined by biases, gender gaps and information manipulation that may arise from the use of AI.

[5] This includes the freedom and honor that may be harmed by threats or defamation using deepfake technology to create fake videos, sexually altered images, or voice impersonations of others.

【Provisional translation】

- Transparency
  - ✧ Appropriately ensuring transparency by disclosing information and securing post-hoc verifiability within necessary and technically feasible limits to enhance reliability in AI[6]
- Accountability[7]
  - ✧ Considering social impact caused by AI, fulfilling accountability within reasonable limits from technical, institutional and social perspective by clarifying where responsibilities lie or constructing initiatives for accountability, etc
- Security
  - ✧ Appropriately ensuring AI security to reduce risks such as AI's unintended operation or stoppage due to unauthorized manipulation
- Privacy
  - ✧ Respecting and appropriately protecting privacy according to the importance of data handled
- Fair competition
  - ✧ Even when resources related to AI are concentrated among specific entities, contributing to promoting fair competition by preventing unfair practices, including the improper collection of data through the exploitation of such advantageous positions
- AI literacy
  - ✧ Acquiring knowledge and abilities as well as maintaining ethical awareness to maximize benefits and minimize risks, recognizing that socially acceptable levels of risks caused by AI may change
- Innovation
  - ✧ Striving to contribute to promote innovation while ensuring sustainability including reducing environmental impact
  - ✧ Engaging in technology development for AI contributing to resolve social challenges
  - ✧ Aiming to improve causes which prevent from utilization of AI

---

[6] It is also important to advance the understanding of AI behavior and the process of generating output from input, thereby deepening our comprehension of the algorithms that contribute to AI output.

[7] It means that individuals and organizations take responsibility for their actions and decisions and take action to fulfill that responsibility.

(3) Basic policy for ensuring appropriateness

Based on the concept described in (2), the following basic policies should be addressed to ensure appropriateness:

- **Risk-based approach**
  Identifying and evaluating risks posed by AI and taking appropriate measures according to the impact based on the field and purpose of AI utilization[8]

- **Active involvement of stakeholders**
  Stakeholders affected by the benefits, risks, and other impacts of AI (hereinafter referred to as "stakeholders")[9] shall actively participate in AI governance, and collaborate with other stakeholders to address challenges.

- **Establishing end-to-end AI governance**
  Managing the risks brought by AI at a level acceptable to stakeholders as well as establishing AI governance that holistically addresses each stage of AI—from research and development to societal implementation—as closely interconnected phases to maximize benefits

- **Agile Response**
  Given the rapid pace of AI technological advancement and the insufficient predictability and explainability, enhancing the maturity of AI governance by responding flexibly and swiftly (hereinafter referred to as "agile") through the PDCA (Plan-Do-Check-Act) cycle to variable risks

---

[8] It is desirable to adopt diverse internal testing methods and independent external testing methods by combining various techniques such as red teaming, and to implement appropriate measures to address identified risks and vulnerabilities.

[9] For example, this may include holders with data for the foundation of AI innovation, those who handle AI outputs and stakeholders who are affected by AI utilization but are not directly involved.

2. Matters to be especially addressed by R&D institutions and utilization business operators

Utilization business operators[10] who develop and provide products and services utilizing AI should, considering that the AI they develop and provide may affect many stakeholders, utilize international norms[11], international standards[12], and various other domestic guidelines[13] related to R&D and utilization of AI, and especially address the following matters in particular regarding the main elements necessary for ensuring appropriateness as indicated in Section 1(2).

Furthermore, when providing the developed AI to third parties, AI R&D institutions[14] shall address the following matters in particular regarding the main elements necessary for ensuring appropriateness as indicated in Section1(2).

(1) Ensuring overall appropriateness through AI governance

Establishing[15] operating, and continuously improving[16] AI governance—including organizational processes for identifying, assessing, and addressing risks throughout the entire AI lifecycle (design,

---

[10] It refers to the utilization business operator as defined in Article 7 of the AI Act, including overseas business operators.

[11] For example: Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems, and Hiroshima AI Process International Guiding Principles for All AI Actors

[12] For example: AI management system (ISO/IEC 42001)

[13] It Refers to the Cabinet Office website (to be updated with domestic and international AI standards, guideline, etc.)

[14] This refers to research and development institutions as defined in Article 6 of the AI Act. Regarding universities among such research and development institutions, consideration shall be given to respecting the autonomy of researchers and other characteristics of university research, as stipulated in Article 6, Paragraph 2 of the AI Act.

[15] It is useful not only to build AI governance from scratch but also to leverage governance processes already applied to existing IT systems and other areas.

[16] At this juncture, to build a highly reliable organization, including measures to avoid risks related to AI and how to respond if risks become apparent, it is considered possible to manage and control AI-related risks and fulfill social and ethical responsibilities by utilizing the “Hiroshima AI Process” reporting framework and establishing and operating a management system based on international standards (such as the AI management system (ISO/IEC 42001)). Proactively disclosing and explaining these initiatives is expected to enhance corporate value and secure competitive advantage.

development, provision, implementation, etc.), such as mechanisms for monitoring and evaluation involving management, appropriate disclosure of related information, and implementation of education and training, and managing AI risks at an acceptable level while maximizing the benefits AI brings.

(2) Ensuring transparency to build trust with stakeholders

Ensuring explainability within reasonable limits[17] to building trust with stakeholders[18] regarding the origin of training data and generated outputs, including the appropriate implementation of intellectual property and privacy protections.

Furthermore, when providing AI, supplying users with information enabling its proper use (such as the AI's mechanisms and limitations, prohibited actions, data collection policies for training, and warnings regarding output reliability[19]).

(3) Ensuring sufficient safety

Identifying and evaluating risks of illegal acts such as cyberattacks and fraud that misuse AI, as well as various other crimes, and taking appropriate measures.

Also, utilizing the latest technologies and knowledge to address and improve issues to suppress inappropriate outputs by AI such as hallucination, expansion of bias or discrimination, spread of misinformation or disinformation (including fake videos by deepfake technology and sexual altered images), and to prevent unintended operations or malfunctions of AI. In particular, given that the spread of misinformation and disinformation generated by AI poses a serious risk, R&D institutions

---

[17] While taking care to protect the trade secrets and intellectual property rights of businesses and other entities, care should be taken not to impose excessive burdens or request excessive disclosure of information.

[18] To ensure appropriate transparency of training data, R&D institutions and utilization business operators will display the information (such as websites) used as the basis for AI outputs, and when disclosure of training data is requested, they will assess the necessity and respond appropriately.

[19] This includes warnings to prevent inappropriate user behavior that could lead to unfair bias or discrimination in hiring, performance evaluations, etc., or the spread of misinformation and disinformation, as well as contact points and contact information for handling user inquiries.

and utilization business operators will strive to develop technologies that can determine whether something is generated by AI (digital watermarks, APIs[20], etc.) and will implement them as necessary.

(4) Maintaining a safe environment through business continuity

Operators of AI-based systems and service providers shall establish in advance a business continuity plan which shall define activities to be performed during normal operations, as well as methods and means for business continuity during emergencies, to minimize damage and enable early recovery in the event of system failures.

(5) Consideration for stakeholders based on the importance of data as the foundation of AI innovation

For AI innovation, it is important to secure high-quality data and use it appropriately. Based on this, in order to realize a virtuous cycle in which new creative activities are promoted by enriching high-quality data and developing and providing reliable AI, businesses developing and providing AI will strive to continuously communicate with stakeholders, such as data holders of intellectual property, about the state of appropriate utilization, depending on the data usage situation. Furthermore, in particular, businesses that develop and provide AI with significant social impact shall endeavor to consider and implement measures aimed at establishing an ecosystem for returning benefits to such as date holders of intellectual property, and creating an environment where creative activities can be conducted with peace of mind.

---

[20] Application Programming Interface: a mechanism for linking different applications (software) and systems to enable communication and data exchange.

【Provisional translation】

3. Matters to be especially addressed by national and local governments

The national government should especially address the following matters in particular regarding the main elements necessary for ensuring appropriateness as indicated in Section 1(2).

Local governments should, considering the diversity of their environments and challenges, respond as necessary with particular attention to the following matters in particular according to local circumstances, regarding the main elements necessary for ensuring appropriateness as indicated in Section 1(2).

When developing and providing AI, the local governments also address the matters indicated in Section 2.

(1) Promoting innovation through active and leading utilization of AI

Recognizing that disseminating practical use cases and key considerations is effective for promoting AI adoption, national and local governments will proactively lead the way in advancing AI utilization and providing development and demonstration opportunities through public procurement.

(2) Improving AI literacy throughout the entire society

The national and local governments are required to promote the enhancement of AI literacy throughout society so that all entities—including, of course, national and local government employees—can understand issues related to ethics, laws, human rights, safety, and so on, and act with an awareness of their responsibilities as users.

To this end, the government will continuously monitor the latest technological trends and practical applications of AI, examine associated risks and countermeasures, and present a concept to encourage voluntary initiatives by stakeholders. Furthermore, to ensure appropriate AI R&D and utilization by businesses and citizens, the government will actively promote education and guidance, including provision of contents that teach the basic usage and precautions for generative AI, and support for working adults in acquiring generative AI skills and knowledge.

(3) Examining appropriate approach to AI governance

The government will closely monitor domestic and international trends in AI governance, continuously review the state of AI governance, and respond accordingly. This guideline and various other domestic guidelines related to R&D and utilization of AI will be reviewed continuously and in an agile manner in order to reflect societal changes driven by technological advances in AI. In doing so, various other

domestic guidelines related to R&D and utilization of AI shall be reviewed and revised as appropriate to be consistent with the intent of this guideline and are easily understandable for businesses, citizens, and others.

Furthermore, to reduce barriers to AI adoption in various contexts, with regard to issues that can be anticipated or that may arise when utilizing AI, the government will organize the issues and ideas regarding the interpretation and application of where responsibility lies, and strive to clarify interpretations as much as possible based on precedents, etc.

Additionally, as AI operates across borders, international governance is essential alongside domestic efforts, the government will take the lead in establishing AI governance while also considering the need to ensure interoperability.

(4) Fulfilling accountability as an administration

When utilizing AI in government administration, to ensure the reliability of government administration, the government will implement appropriate risk countermeasures that fully consider the required standards, and fulfill our accountability to the public by ensuring that the basis for decisions remains as clear as possible.

Moreover, each ministry and agency shall appoint an AI governance officer[21]. Local governments shall clearly designate officers responsible for the appropriate utilization of AI and risk management.

---

[21] "The Guideline for Japanese Governments' Procurements and Utilizations of Generative AI for the sake of Evolution and Innovation of Public Administration" (Approved by the Council for the Promotion of a Digital Society Executive Board Meeting on May 27, 2025) stipulates that each ministry and agency shall establish and promote policies for the utilization of generative AI, and shall establish a Chief AI Officer (CAIO) to oversee the overall utilization status and risk management across the organization. When utilizing AI other than generative AI, it is also desirable to clearly designate responsible personnel as necessary.

【Provisional translation】

4. Matters to be especially addressed by citizens

Citizens shall respond with particular attention to the following matters regarding the main elements necessary for ensuring appropriateness as indicated in Section 1(2).

(1) Responsible use of AI based on principles of human-centric AI

Citizens, as the primary users of AI, shall comply with laws and regulations, recognizing that AI use may lead to violations of laws or harmful actions.

In addition, citizens shall strive to understand not only the convenience of AI but also issues concerning ethics, laws, human rights, and safety, acting with awareness as responsible users.

(2) Appropriate use based on AI literacy

Citizens strive to correctly understand the characteristics and mechanisms of AI and proactively acquire AI literacy.

Furthermore, when utilizing AI, citizens shall understand the source and accuracy of the information obtained, make decisions under human judgment and responsibility, and refrain from inappropriate actions aimed at unfair bias or discrimination, slander, and spread of disinformation and misinformation. Additionally, citizens shall utilize AI-generated outputs (text, images, audio, video, etc.) in socially and legally appropriate ways.