仮訳
Provisional Translation

AI Guidelines for Business Ver1.2

March 31, 2026

Ministry of Internal Affairs and Communications
Ministry of Economy, Trade and Industry

Contents

Preface 2
Part 1 Definitions 9
Part 2 Society to aim for with AI, and matters each AI business actor works on 12
A. Basic philosophies 12
B. Principles 13
C. Common Guiding Principles 14
D. Hiroshima AI Process “Hiroshima Process International Guiding Principles for All AI Actors” 25
E. Building AI governance 26
Part 3 Matters Related to AI Developers 29
Part 4 Matters Related to AI Providers 37
Part 5 Matters Related to AI Business Users 40

Preface

Technologies related to AI (Artificial Intelligence) are constantly evolving, opportunities to use AI and its various possibilities have been increasing continuously, and AI is being used for making industrial innovations and solving social challenges as well. In addition, interactive generative AI has brought about the "democratization of AI," enabling many people to easily utilize AI for a wide variety of purposes through "dialogue." Therefore, companies have been making efforts to not only incorporate AI into their business processes but also reconstruct their business models themselves based on the values that will be created by AI. Some private individuals have also increased their efforts to apply their knowledge to AI and enhance their productivity. Japan has been promoting Society 5.0, a concept of a human-centric society in which both economic growth and solutions for social challenges are achieved through a system that merges the cyberspace and physical space in an advanced way, called a "Cyber-Physical System" (CPS). To embody this concept and enable the society to accept AI and use it appropriately, "Social Principles of Human-Centric AI" was established in March 2019. Meanwhile, risks have been increasing as the scope of use of AI technologies and the users have been increasing. Generative AI, especially, has incurred new societal risks that are not carried by conventional AI, such as infringements of intellectual property rights and generation and transmission of disinformation or misinformation, leading to the diversification and increase of societal risks resulting from AI. In light of these circumstances, the "Act on Promotion of Research and Development, and Utilization of AI-related Technology" (Act No. 53 of 2025) was promulgated in June 2025 and came into full effect in September of the same year.[1, 2]

These Guidelines present unified guiding principles for AI governance in Japan to promote the safe and secure utilization of AI. It is intended to help people who use AI in various businesses to fully recognize AI risks based on international trends and stakeholders' concerns, and to voluntarily take the necessary countermeasures across the entire lifecycle. The Guidelines aim to actively and cooperatively develop a framework that achieves both promotion of innovation and reduction of risks across the lifecycle through mutual cooperation among interested parties in implementing the common guiding principles, important matters for each AI business actor, and AI governance.

Japan led discussions held at international forums, such as the G7, G20, and OECD, and made a lot of contributions, starting with the proposal for the AI R&D Principles at the G7 ICT Ministers' Meeting in Takamatsu, Kagawa, in April 2016. Incidentally, the following matters have been pointed out regarding the actual implementation of the principles in AI:

- AI use is viewed as a solution to some social challenges, such as decreasing labor caused by a declining birthrate and aging population.
- There is a time lag between formulation and enforcement of laws and the speed and complexity of AI technology development and social implementation.
- Rule-based regulations that stipulate detailed obligations might inhibit innovations.

Thus, it was decided to draw up guidelines on the basis of the goal-based concept that would lead to the achievement of purposes through soft laws without any legally binding force that would encourage interested parties to make voluntary efforts to reduce societal risks in AI and promote innovations and use of AI.

On this understanding, the "Draft AI R&D Guidelines for International Discussions" and "AI Utilization Guidelines: Practical Reference for AI Utilization" were established and announced on

[1] https://www8.cao.go.jp/cstp/ai/ai_act/ai_act.html
[2] In December of the same year, pursuant to Article 13 of the Act, the "Guideline for Ensuring the Appropriateness of Research & Development and Utilization of Artificial Intelligence-Related Technology" (decided by AI Strategic Headquarters on December 19, 2025) were formulated to encourage voluntary and proactive efforts by all AI-related actors for the appropriate conduct of AI research, development, and utilization.
https://www8.cao.go.jp/cstp/ai/ai_guideline/ai_guideline.html

the initiative of the Ministry of Internal Affairs and Communications, and "Governance Guidelines for Implementation of AI Principles ver. 1.1" were established and announced on the initiative of the Ministry of Economy, Trade and Industry. And by integrating and reviewing those three guidelines, the Guidelines (non-binding soft law) were established in April 2024 as guidelines to help business operators jointly practice both social implementation of AI and governance thereof, reflecting the features of AI technologies and the domestic and international discussions on social implementation of AI. (See "Figure 1. Positioning of the Guidelines.") It is intended to help business operators (including public institutions such as governments[3] and municipalities[4]) who use AI by referring to the Guidelines instead of the existing guidelines to understand the guiding principles that lead to desirable actions for safe and secure use of AI. The Guidelines are established through studies conducted by multiple stakeholders that consisted of academic and research institutions, civil societies including general consumers, private sector companies, and the like, rather than having the government take the initiative alone, to prioritize effectiveness and validity.

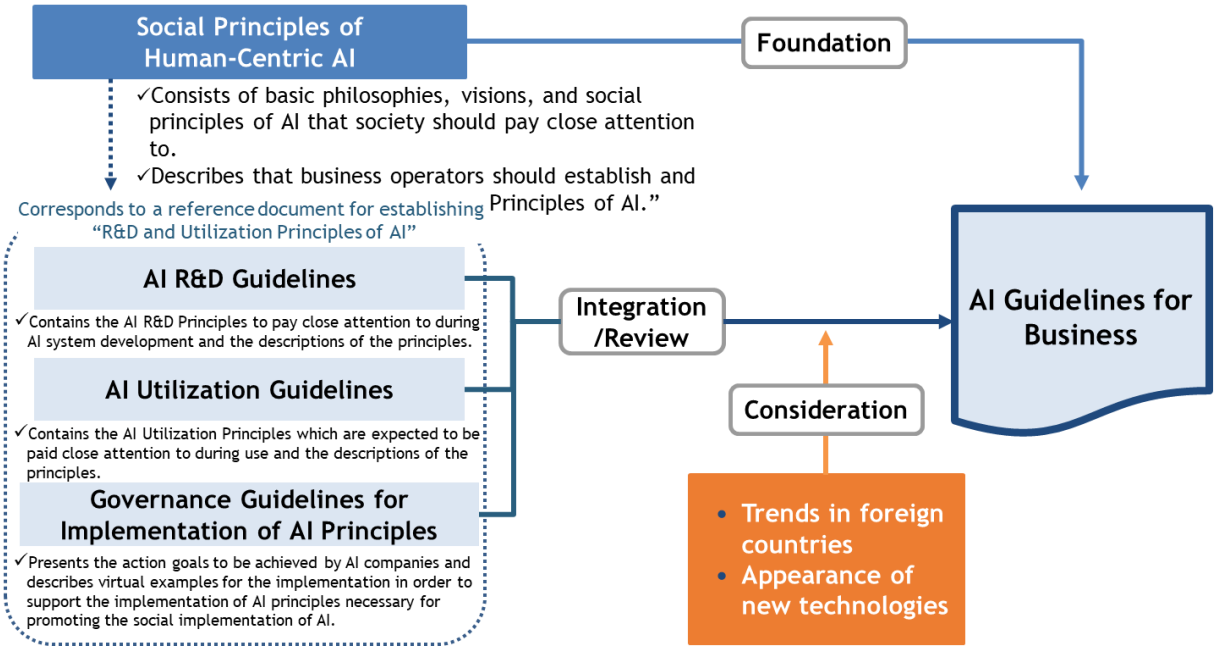

Figure 1. Positioning of the Guidelines

The use of AI might incur a great risk to society depending on the field and how it is used, and social disagreement caused by this risk might inhibit AI use itself. On the other hand, taking too many measures might inhibit AI use itself or decrease the benefits of using AI. Therefore, it's important to take a risk-based approach in which we estimate the level of risks (impact and

[3] The government has formulated the "The Guideline for Japanese Governments' Procurements and Utilizations of Generative AI for the sake of Evolution and Innovation of Public Administration" (decided by the Council for the Promotion of a Digital Society Executive Board Meeting on May 27, 2025) in order to promote the utilization of generative AI in various government operations and to manage associated risks as two sides of the same coin. These Guidelines set out the government's approach to AI governance and frameworks for sharing best practices, risks to be considered in the procurement and utilization of generative AI, and the improvement of the overall functionality, quality, and cost-effectiveness of generative AI utilized by the government. They were developed taking into account existing guidelines such as the AI Guidelines for Business and the "Common Standards Group for Cybersecurity Measures for Government Agencies and Related Agencies," as well as trends in rules adopted by foreign governments, and are presented as guidelines for national government employees and other relevant personnel.
https://www.digital.go.jp/resources/standard_guidelines#ds920

[4] Regarding AI usage methods and considerations for use by municipalities, the "The Handbook on the Use and Introduction of AI in Local Governments <Introduction Procedures Edition>" (revised in December 2025) has been published, which provides content specific to municipalities while referencing these Guidelines. The revised Guidebook particularly highlights the potential for dramatic improvements in operational efficiency expected through the utilization of generative AI, presenting this alongside examples of utilization in municipalities, while also presenting considerations such as establishing structures for ensuring governance, the handling of information requiring confidentiality, and approaches to human resource development.
https://www.soumu.go.jp/main_content/000820109.pdf

probability of hazards) that can be entailed by how it is used in the applicable field, to ensure the degree of measures taken are appropriate to the level of risks. The Guidelines provide the guides to the measures to be taken by companies based on a risk-based approach. Note that the concept of a risk-based approach has been commonly adopted by countries with advanced AI.

Trends in AI are rapidly changing, and in light of international discussions and other developments, it is planned that the Guidelines will be appropriately updated as a living document, including the definitions and other content herein, with multiple stakeholder engagement, while reflecting the agile governance philosophy toward the continuous improvement of AI governance[5]. In such activities, it will be determined how the guiding principles and implementation should be updated as the countermeasures against risks in accordance with maturity of AI in the society. (See "Figure 2. Basic concepts of the Guidelines.")

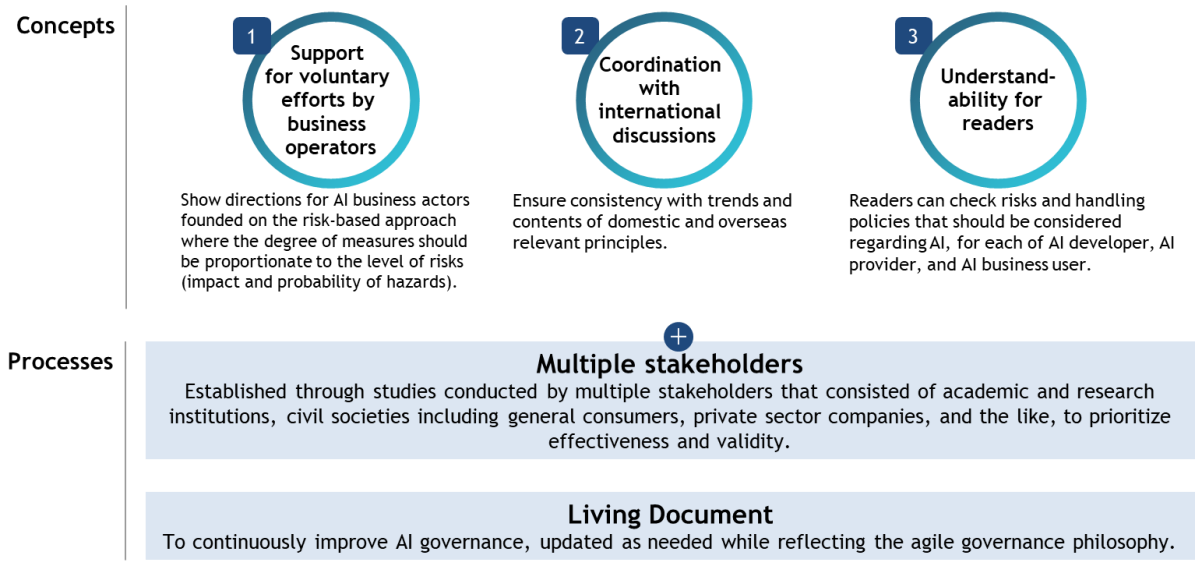

Figure 2. Basic concepts of the Guidelines

The Guidelines present basic concepts regarding efforts necessary for the development, provision, and use of AI. Therefore, for the actual development, provision, and use of AI, it is important that all business operators who intend to use AI will voluntarily promote specific efforts using the Guidelines as one of their references. At the same time, all business operators who intend to use AI should recognize the magnitude of AI's impact on society and be conscious of using it to develop human society better. It is important that the business operators pay close attention that if society considers the efforts to be inappropriate or insufficient, it might lead to opportunity losses in their businesses and it might become difficult to maintain business values. Paying close attention to these possibilities enables to maximize the benefits from AI, strengthen competitiveness, and maintain and improve the business values. Incidentally, because the Guidelines contain reference information for AI use and risk information, it is also helpful for people who are not business operators relevant to AI, for example, staff of academic and research institutions and general consumers (including minors).

The Guidelines are intended for all AI business actors (including public institutions such as governments and municipalities) who develop, provide, or use AI in various businesses. On the other hand, the Guidelines are not intended for those who use AI for non-business activities and those who derive benefits from AI systems and services without directly using AI for business and, in some cases, sustain damage (hereinafter, both are referred to as "non-business users").

[5] The report is compiled by the Ministry of Internal Affairs and Communications' conference toward AI Network Society and the Ministry of Economy, Trade and Industry's Study Group on AI Business Guidelines. The review system will be determined and modified as appropriate in line with future circumstances.

However, necessary points for those who develop, provide, or use AI for business purposes to serve non-business users are included in this guideline. Data is dispensable for AI to learn. Specific companies and individuals (hereinafter referred to as "data providers") who provide such data are similarly not included in the target of this guideline. This guideline assumes those who develop, provide or use AI are themselves responsible for those data as data holders.

As described above, the parties that the Guidelines are intended for are roughly grouped under "AI developers," "AI providers," and "AI business users" as AI business actors who conduct AI businesses and are defined herein. It is assumed that these AI business actors are business operators (or departments of business operators), and a business operator might take on two or more roles as an AI developer, AI provider, and AI business user depending on the AI use method. (See "Figure 3. Correlation between AI business actor and general AI use flow.")[6, 7]

- **AI developer**
  Business operators who develop AI systems (including business operators who research AI). They develop AI models as well as algorithms and contribute to construction of AI systems including AI models, base system, as well as I/O functions via data collection (including purchase), data preprocessing, training with data.[8]
  In addition, even after the development and operational deployment of AI models and systems, they are also responsible for maintaining and improving model performance through post-training (post-training), which aims to expand domain knowledge in specific areas, adapt to changes in the environment, and further make adjustments (alignment) to ensure behaviors aligned with human intentions and values.

- **AI provider**
  Business operators who incorporate AI systems into applications, products, or existing systems, business processes, etc., and provide them to AI business users and, in some cases, non-business users as services.
  They verify AI systems, integrate AI systems with other systems, provide AI systems and services, offer operation support for AI business users on AI systems for normal operations, or perform the AI service operation itself. Communication with various stakeholders might be required during the provision of AI services.

- **AI business user**
  Business operators who use AI systems or AI services in their businesses.
  Their role is to use an AI system or AI service in an appropriate way intended by the AI provider, share information such as environmental changes with the AI provider, continue the normal operation, operate the provided AI system as necessary. In addition, when non-business users might be affected by AI use in some ways[9], AI business users are also responsible for making efforts to prevent AI from incurring unexpected disadvantages for those non-business users and maximize benefits from AI.

[6] The current status, outlook, and challenges and other aspects of digital technologies, including AI, that can be utilized to support consumers are summarized in the report of the "Expert Panel on Digital Technologies Empowering Consumers," published by the Cabinet Office.
https://www.cao.go.jp/consumer/iinkaikouhyou/2024/doc/202412_digital_technology_houkoku.pdf
[7] If an AI provider or AI business user is a public institution, such as a government or municipality, a concept different from that for private business operators might be required.
[8] Generally, AI developers are responsible for API specification formulation, input/output design, and infrastructure development for operating AI models, while AI providers are responsible for UI/UX design and integration with existing business systems. Therefore, it is not the case that AI developers are responsible for all aspects of "building AI systems." Moreover, in practice, diverse cases exist, and responsibilities are not limited to those described above.
[9] Non-business users need to pay close attention that they may suffer some type of damage if they do not follow the instructions and precautions from AI business users.

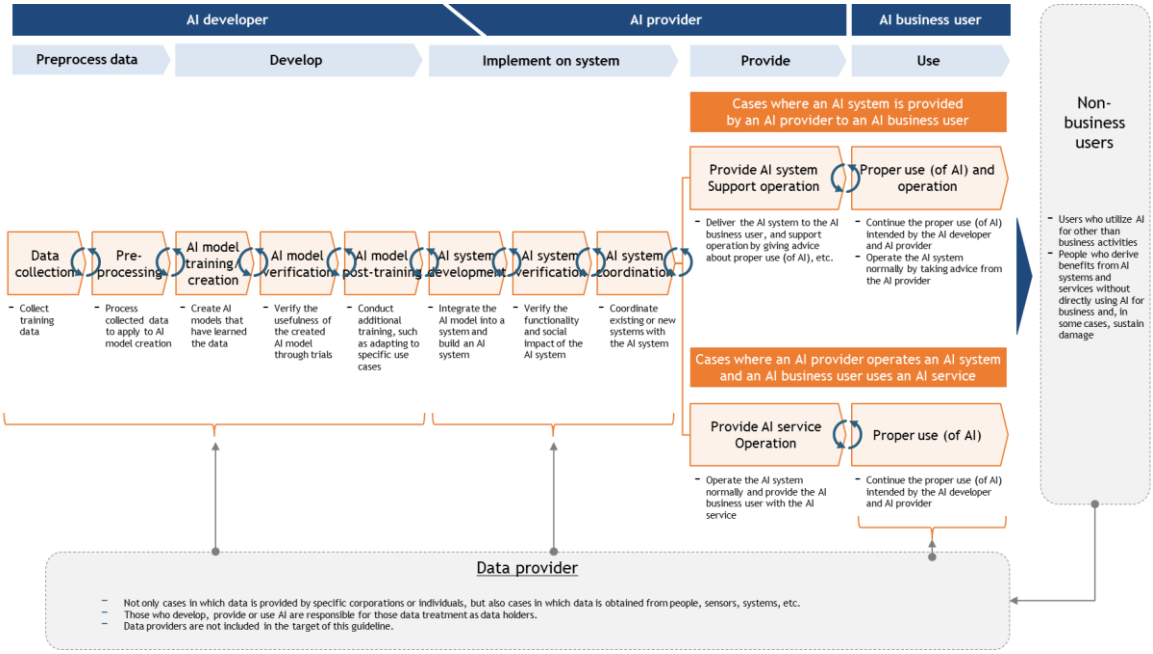

Figure 3. Correlation between AI business actor and general AI use flow

It is important that each party clarifies “the efforts to be made regarding AI (guiding principles = what)” based on “the ideal society while considering stakeholders’ expectations (basic philosophies = why)” from each relevant perspective (AI developer, AI provider, or AI business user). In addition, it is conceivable that studying, determining, and implementing “the specific approach to be adopted (implementation = how)” to fulfill the guiding principles are useful to use AI safely and securely. Actual AI systems and services can be used in various cases depending on the purpose, used technology, data, usage environment, etc. Therefore, it is important that AI developers, AI providers, and AI business users cooperate with each other to devise the optimum approach while considering changes in the external environment, such as the advancement of technologies. For the sake of readability, the main part of the Guidelines covers the basic philosophies and guiding principles, and the appendix covers implementation.

The structure of the main part of the Guidelines, which covers the basic philosophies and guiding principles, is shown below:

- **Part 1**
  This part mainly describes definitions of terms to help understand the Guidelines.

- **Part 2**
  This part describes the society to aim for through AI use, the basic philosophies (why) and principles for realizing it, and the common guiding principles (what) among AI business actors. It also describes the establishment of governance required for implementing the common guiding principles considering the possibility of risks in AI to the society during the pursuit of benefits from AI use. Part 2 describes matters that form the base for Part 3 and later parts, so it is important that all business operators who use AI read it and understand its descriptions.

- **Parts 3 to 5**
  These parts describe the precautions for each of the three AI business actors who conduct businesses using AI that are not mentioned in Part 2. It is important that each business operator who uses AI understands the precautions relevant to itself. In addition, it is also important that each AI business actor understands the precautions for other AI business

actors as well, because there are many matters relevant to adjacent AI business actors. (See “Figure 4. Structure of the Guidelines.”)

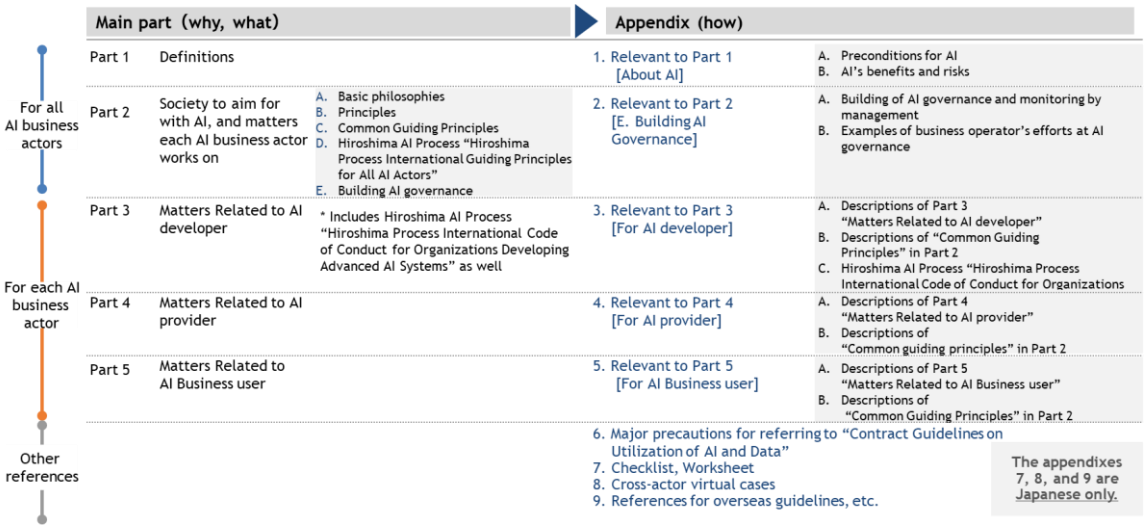

Figure 4. Structure of the Guidelines

For AI developers, AI providers, and AI business users, in addition to Parts 1 and 2, reading the corresponding part among Parts 3 to 5 and the appendix will help them understand risks in AI use and the basic concept of the policy for handling them. Because the examples described in the appendix are helpful as references for business operators who have not determined any specific efforts to make, it is important to chiefly read relevant contents in the appendix. For business executive officers[10], including management, to fulfill their duties, it is important to consider and take countermeasures against risks in AI use together with the business strategy, in accordance with the basic philosophies (why) and guiding principles (what) described in the Guidelines to promote safe and secure AI use.

The environment surrounding AI is rapidly advancing worldwide, so it is important that business operators who intend to use AI pay attention to international trends. Under these circumstances, Japan took the initiative in establishing an international common understanding on AI and its guiding principles through the Hiroshima AI Process[11] and took on the key role in the development of the Hiroshima AI Process Comprehensive Policy Framework in December 2023.[12] The Guidelines

[10] Business executive officers include those in public institutions such as governments and municipalities.

[11] Based on the result of the G7 Hiroshima Summit held in May 2023, the Hiroshima AI Process was initiated to study international rules concerning generative AI. After that, the G7 Digital & Tech Minister Meeting was held in December 2023 based on the “G7 Leaders' Statement on the Hiroshima AI Process” announced after the Hiroshima AI Process Minister-level Meeting in September 2023 and the Multi-stakeholder High-level Meeting at Kyoto IGF in October. As the achievements of 2023, the “Hiroshima AI Process Comprehensive Policy Framework” was formulated. Furthermore, in 2024, following the “G7 Ministers’ Meeting on Industry, Technology, and Digital” in March and the “G7 Digital and Technology Ministers’ Meeting” in October, a “reporting framework” concerning adherence to international codes of conduct was agreed upon by the G7 in December of the same year. https://www.soumu.go.jp/hiroshimaaiprocess/Note that, as of March 2026, 25 organizations, including 9 Japanese companies, have submitted their responses, which are published on the OECD website.
https://transparency.oecd.ai/reports

[12] GPAI (the Global Partnership on Artificial Intelligence) was established in 2020 as a multi-stakeholder international collaboration initiative involving governments, international organizations, industry, experts, civil society, and other stakeholders, to promote the development and use of “responsible AI” through project-based efforts, based on a human-centered approach. In July 2024, GPAI transitioned to an integrated partnership with the OECD. In order to provide support in operations and management, including projects related to generative AI, which also aid in advancing the Hiroshima AI Process, the “GPAI Tokyo Expert Support Center” was established at the National Institute of Information and Communications Technology (NICT) in July 2024. The Center collaborates with the OECD Secretariat and other GPAI Centers (Inria (France) and CEIMIA (Canada)) and promotes various projects, including research and analysis on the latest AI trends (such as agentic AI), fostering young talent (such as the Student Community), and hosting expert workshops (such as the Innovation Workshop).

are also intended to contribute to the Process and have been established taking into account international discussions including the Process. On the other hand, the policies and rules for AI vary with country and region, so business operators who perform cross-border activities should obey local laws and fulfill stakeholders' expectations. As for advanced AI systems, especially, some countries and regions take some measures to assure effective AI governance, for example, establishing a safety framework of AI validation prior to their release to the market[13, 14], so it is important to pay attention to them[15].

https://www2.nict.go.jp/gpai-tokyo-esc/
https://www2.nict.go.jp/gpai-tokyo-esc/agenticAI/

[13] In November 2023, the UK announced a plan for founding the AI Safety Institute that would develop and perform evaluations of advanced AI systems. The US announced that it would establish the US AI Safety Institute in the National Institute of Standards and Technology (NIST) to implement an AI risk management framework and evaluate red teaming. In the UK, the AI Safety Institute was renamed the AI Security Institute in February 2025. In Japan, in collaboration with these overseas institutions, the "AI Safety Institute" (AISI) was established within the Information-technology Promotion Agency (IPA), an incorporated administrative agency, on February 14, 2024, with the cooperation of relevant government ministries, agencies, and related organizations, for the purposes of considering standards and guidance contributing to the improvement of safety in AI development, provision, and use, conducting surveys on AI safety evaluation methods, and investigating technologies and cases related to AI safety. Its activities include developing a cross-walk between this guideline and the US NIST AI Risk Management Framework (RMF), publishing the "Guide to Evaluation Perspectives on AI Safety (Version 1.10)" (March 2025) https://aisi.go.jp/output/output_information/250328_1/, releasing the AI Safety Evaluation Tool, collaborating and exchanging views with AISIs and other counterparts in various countries, and conducting various surveys.
https://aisi.go.jp/effort/effort_framework/guide_to_evaluation_perspective_on_ai_safety/https://aisi.go.jp/output/output_information/250912/
https://aisi.go.jp/
[14] For example, as an initiative toward ensuring the reliability and safety of AI, the National Institute of Information and Communications Technology (NICT) is advancing research and development toward the building of an "active evaluation platform" for evaluating the reliability and safety of Large Language Models (LLM).
[15] On February 4, 2025, the interim report of the AI Strategy Council and AI Institutional Research Group was released. (Japanese Only) https://www8.cao.go.jp/cstp/ai/interim_report.pdf

Part 1 Definitions

The term “AI,” which means Artificial Intelligence, is said to first used at the Dartmouth Conference in 1956 for the first time. Although there is no agreed definition of AI, as implied from the fact that it is the abbreviation of “Artificial” and “Intelligence,” it refers to a computer program that works in a similar way to human’s thinking process and a system that can make intelligent decisions on a computer. In the past, some systems called “expert systems,” which make inferences from a large amount of knowledge data that is input based on experts' knowledge without machine learning (ML), were considered as a type of AI.

However, in 2000s and later, deep learning emerged, and they used it for image recognition, natural language processing (including translation), and speech recognition with machine learning. And systems able to predict, propose or make decisions in specific areas are now called AI. In addition, since 2021, the rise of foundation models[16] has driven the development of general-purpose AI that is not limited to specific fields. As a result, in addition to “prediction,” “recommendation,” and “decision-making,” “generative AI,” which generates images, text, and other content, has also been gaining widespread adoption. Moreover, in recent years, research and attempts at social implementation of AI agents and physical AI have also been progressing. As described above, AI encompasses a wide variety of types, and it is difficult for even experts to predict the future direction of AI technologies.

With these circumstances, related terms in the Guidelines are defined as follows.

Related terms

- **AI**
  No agreed definition has been existed as of now (“Social Principles of Human-Centric AI” formulated by the Integrated Innovation Strategy Promotion Council on March 29, 2019), and it is difficult to strictly define artificial intelligence in a broad sense. AI in the Guidelines is an abstract concept, which includes AI systems (defined below) themselves or software or programs that perform machine learning.

- **AI system**
  A system (such as a machine, robot, and cloud system) that works at various levels of autonomy during the use process and incorporates a software element that has a learning function.
  (For reference, it is defined in JIS X 22989:2023 based on ISO/IEC 22989:2022 as follows.)
  An engineering system that produces outputs such as contents, predictions, recommendations, and decision-makings in response to a given set of goals defined by humans.
  Note 1: As for an engineering system, models that represent data, knowledge, processes, etc., that can be used to perform tasks can be developed using various techniques and approaches relevant to artificial intelligence.
  Note 2: An AI system is designed to work at various autonomous levels.
  (For reference, it is defined in the OECD AI Principles overview as follows.)
  An AI system is a machine-based system that, for explicit or implicit objectives, makes inferences. It generates outputs including predictions, contents, recommendations, decisions and so on to place impact on physical or virtual environments from received data. Different AI systems vary in their levels of autonomy and adaptiveness after deployment.

[16] Fundamental models, typified by large language models, are core technological foundations for creating individual models that support various services. They have characteristics different from conventional AI in terms of development of models adapted to a broad range of downstream tasks derived from fundamental models, as well as knowledge acquired through the development process itself.

- **Advanced AI system**
  The most advanced AI systems including the cutting-edge foundation models and generative AI systems.
  (Quoted from the definition in the Hiroshima AI Process)

- **AI model (ML model)**
  A model incorporated into an AI system and acquired through machine learning using training data. It produces prediction results in accordance with the input data.
  (For reference, it is defined in JIS X 22989:2023 based on ISO/IEC 22989:2022 as follows.)
  A mathematical structure that produces inferences or predictions based on input data or information.
  Example: When a univariate linear function $y = \theta_0 + \theta_1 x$ is trained using the linear regression, the result model is $y = 3 + 7x$ or the like.
  Note 1: A machine learning model is acquired as a result of training based on a machine learning algorithm.

- **AI service**
  A service that uses AI systems. It refers to providing AI business users with values in general. AI services are provided and operated through not only technologies that constitute an AI system, but also non-technological approaches, including monitoring by humans and appropriate communications with stakeholders.

- **Generative AI**
  A general term representing AI developed from an AI model that can generate texts, images, programs, etc.

- **AI governance**
  The design and operation of technological, organizational, and social systems by stakeholders for the purpose of managing risks posed by the use of AI at levels acceptable to stakeholders and maximizing their positive impact (benefit).

- **Training**
  Training is the process of determining or improving the parameters of an AI model using data, and consists of two processes: pre-training, which builds the foundation model, and post-training[17] , which involves continuous parameter adjustments as needed. Training encompasses methods such as supervised learning, unsupervised learning, and reinforcement learning. The data used for training is divided into training data, validation data, and test data, which are used for the building and evaluation of AI models.

- **Inference**
  Inference is the process of providing new data to a trained model and computing outputs (such as predictions, classifications, and generation). In the inference process, data for inference is used, which includes information that the model directly processes, such as prompts from users and sensor-acquired data, as well as additional information referenced through mechanisms such as RAG (Retrieval-Augmented Generation) as needed.

- **AI Agent[18]**

[17] Post-training refers to efforts to continuously adjust the parameters of a pre-trained AI model as needed, thereby expanding domain knowledge in specific areas, achieving control (alignment) so that the model behaves in alignment with human intentions and values, and enabling adaptation to changes in the environment. Examples of post-training include fine-tuning to obtain performance suited to specific operations, reinforcement learning from human feedback (RLHF), and retraining with new data.
[18] Agentic AI is a concept that is more comprehensive and evolutionary than AI agents. It is a goal-driven AI system that autonomously makes decisions and takes actions through the use of multiple AI agents.

In these Guidelines, an AI agent refers to an AI system that perceives its environment and acts autonomously[19] to achieve a specific goal.

(For reference, ISO/IEC 22989:2022 provides the following definition:)

An automated entity that senses and responds to its environment and takes actions to achieve its own goals.

- **Physical AI**

  In these Guidelines, physical AI refers to a system that takes in environmental information through sensors, processes that information using an AI model, autonomously infers and determines strategies for achieving objectives given by humans, and furthermore acts on those strategies without human intervention.

[19] "Autonomy" here does not refer solely to a highly autonomous state, but also includes systems with a certain degree of autonomy.

Part 2 Society to aim for with AI, and matters each AI business actor works on

Part 2 describes “A. Basic philosophies” as the society to aim for with AI first. Next, it describes “B. Principles” at which each AI business actor works on to realize the basic philosophies and “C. Common Guiding Principles” that are derived from the principles. Furthermore, it describes “D. Hiroshima AI Process 'Hiroshima Process International Guiding Principles for All AI Actors' ,” established in the Hiroshima AI Process. After that, it describes “E. Building AI governance,” which is important for the implementation of “C. Common guiding principles” and safe and secure use of AI.

A. Basic philosophies

As described in the “Preface,” “Social Principles of Human-Centric AI” formulated by Japan in March 2019 states that it is expected that AI will contribute to the creation of Society 5.0. Additionally, the document states that it is important to use AI as a public asset of humans that can lead to the achievement of global sustainability through qualitative changes of the ideal society as well as true innovations. The document also states that the following three values should be respected as basic philosophies in order to build a society that upholds such philosophies.

(1) **Dignity: A society that has respect for human dignity**
Rather than building a society in which effectiveness and convenience are pursued through AI use to the point that humans become excessively dependent on AI and AI is used to control human behaviors, it is necessary to build a society that has respect for human dignity and for humans to take full advantage of AI as a tool to fully demonstrate their various capabilities. This will allow them to exert more creativity or engage in more challenging jobs and live physically and mentally rich lives.

(2) **Diversity & Inclusion: A society where people with diverse backgrounds can pursue their own well-being**
One of the present-day ideals and big challenges is the creation of a society in which people with diverse backgrounds, values, or ways of thinking can seek different well-being, so they can be flexibly included and new values can be created. Powerful AI technologies can be an effective tool for approaching this ideal. We need to transform the state of society as described above through proper development and deployment of AI.

(3) **Sustainability: A sustainable society**
We need to use AI to bring new businesses and solutions into the world one after another to build a sustainable society that can eliminate social disparities and address global environmental problems and climate changes. As a science-and-technology-oriented country, Japan has a responsibility to contribute to the creation of such a society by using AI to strengthen its accumulated scientific and technological expertise and knowledge.

Figure 5. Basic philosophies

These fundamental concepts remain goals for us to achieve and do not change despite significant technological evolution. Therefore, these basic philosophies should be respected as objectives to achieve through domestic and international frameworks as AI evolves.

B. Principles

To realize the basic philosophies, it is important that each AI business actor makes efforts to comply with the philosophies. Therefore, we categorized the principles that should be kept in mind by each AI business actor into the activities to be implemented by each AI business actor and the activities expected to be implemented in cooperation with society. These principles have been formulated by restructuring "Social Principles of Human-Centric AI" in accordance with overseas principles, including OECD's AI principles.

Activities to be implemented by each AI business actor

It is important that each AI business actor achieves its purposes of AI, such as creating business value and solving social challenges, while promoting the development, provision, or use of AI systems and services and maintaining human dignity, based on the human-centric[20] concept derived from the basic philosophies. To accomplish this, it is important that each AI business actor ensures values, such as safety and fairness, to reduce societal risks arising from AI use. In addition, it is important to protect privacy including the prevention of inappropriate use of personal data and ensure security against risks such as a decreasing availability and external attack caused by vulnerabilities of AI systems. To achieve these goals, it is important that each AI business actor ensures the verifiability of systems and improves transparency by providing appropriate information to stakeholders[21] and ensures accountability.

In addition, it is important that, taking into account the possibility that the roles of each actor may shift due to value chain fluctuations arising from the diversification of AI architectures, among other factors, actors across the value chain collaborate with one another to strive to improve the overall quality of AI throughout the value chain, and that multi-stakeholder discussions continue on an ongoing basis.

By making these efforts, each AI business actor is expected to derive maximum benefit from the development, provision, or use of AI systems and services while minimizing the AI risks.

[20] The underlined parts are organized as "C. Common Guiding Principles" in the latter part.
[21] Stakeholder: All the AI business actors who might be directly or indirectly affected by AI use including third parties other than AI developers, AI providers, AI business users, and non-business users. (The same shall apply hereafter.)

Activities expected to be implemented in cooperation with society

In order to enhance benefits from AI for the society and realize the basic philosophies that we should pursue, each AI business actor is expected to actively collaborate with the society, including the governments, municipalities, and communities, as well as individually commit to its own activities. To accomplish this, each AI business actor is expected to provide opportunities for ensuring education and literacy in cooperation with the society to avoid divisions within the society and spread the benefits from AI to all of the people. In addition to that, each AI business actor is expected to contribute to activities that ensure fair competition and facilitate innovation that can create new businesses and services, maintain sustainable economic growth, and provide solutions for social challenges.

C. Common Guiding Principles

In the activities, each AI business actor should develop, provide, or use AI systems and services respecting the rule of law, human rights, democracy, diversity, inclusion, and fair and just society in light of "1) Human-centric" described below. In addition, relevant laws, including the Constitution of Japan, Intellectual Property Basic Act and relevant laws, and Act on the Protection of Personal Information as well as existing laws and regulations in individual fields pertaining to AI should be observed, and it is important to pay close attention to the circumstances of the drafting of international guiding principles[22] [16].

It is important that each AI business actors understand characteristics, intended use, purposes and social context of AI systems/services and positively process these activities with limited resources.

1) Human-Centric

When developing, providing, or using an AI system or service, each AI business actor should act in a way that does not violate the human rights guaranteed by the Constitution of Japan or granted internationally, as the foundation for accomplishing all matters to be conducted, including the matters described later. In addition, it is important that each AI business actor acts so that the AI expands human abilities and enables diverse people to seek diverse well-being.

(1) Human dignity and autonomy of individuals

- ✧ Based on the social context of AI use, respect human dignity and the autonomy of individuals.
- ✧ In particular, when linking AI with someone's brain or body, refer to discussions about bioethics in foreign countries and research institutions, together with information of peripheral technologies.
- ✧ When profiling using AI in a field where personal rights and benefits can be severely affected, use AI respecting the dignity of individuals, maintaining the utmost accuracy of the outputs, understanding limitations of predictions, recommendations, and judgments of AI, and carefully considering possible drawbacks, and do not use it for inappropriate purposes.

(2) Paying attention to manipulations by AI on decision-makings and emotions

- ✧ Do not develop, provide, or use AI systems and services with purpose of manipulating human decision making, recognition and emotion or on the premise of unconscious control.

[22] Governing laws should be obeyed in accordance with the geographic business deployment status, locations of the AI providers and AI business users of the developed AI models, the locations of the servers to be used for training and the like. When you are required to comply with Japanese laws, handle personal data, intellectual property rights, and such like in compliance with their respective applicable laws in accordance with the data type. As for the handling of data, note that the use of some data might be prohibited by a contract between stakeholders, even when it is not stipulated by laws.

- ✧ When developing, providing, or using an AI system or service, pay attention and take necessary countermeasures against the risk of heavy dependence on AI, such as automated biases[23] [24].
- ✧ Pay attention to AI use that might instigate biased information or values and unwillingly limit the options that should be originally available to people including AI business users, such as a filter bubble[25].
- ✧ Carefully handle AI outputs, especially when they can be relevant to procedures that might significantly affect the society, such as an election and decision-making in a community.

(3) Countermeasures against disinformation, etc.

- ✧ Generative AI has enabled everyone to forge fake information that seems to be true and fair, so recognize the increasing risk of destabilizing and confusing the society through disinformation, misinformation, and biased information generated by AI, and take necessary countermeasures[26, 27].

(4) Ensuring diversity/inclusion

- ✧ In addition to ensuring fairness, to prevent information poverty and digital poverty and allow more people to enjoy the benefits of AI, pay attention to make it easy for socially vulnerable people to use AI.
  - Adopt universal design, ensure accessibility, and provide relevant stakeholders[28] with education and support.

(5) Providing user support

- ✧ Offer rational information about the functions and peripheral technologies of the AI system or service, and allow users to use functions that timely and appropriately offer the information for judging choices.
  - For example, default settings, provision of understandable options, provision of feedbacks, alerts in an emergency, and handling of errors.

(6) Ensuring sustainability

- ✧ Examine the impact of the whole lifecycle on the global environment during the development, provision, and use of AI systems and services. (In particular, for AI systems with high computational complexity, such as generative AI, implement

[23] Refers to a phenomenon in which automated systems or technologies are excessively trusted or depended on when humans make judgments and decisions.

[24] The necessary measures proposed in response to the above challenges include the following methods:

・Receiving training to understand the characteristics and limitations of AI, in order not to blindly accept AI outputs, such as evaluations, judgments, recommendations, or predictions.

・When approving AI evaluations or judgments, independently considering the reasons and grounds for their approval before doing so.

[25] A filter bubble refers to an information environment where an algorithm analyzes and learns about search histories and click histories of individual Internet users to preferentially show information they like, regardless of whether or not they want it to do so, separating them from information that disagrees with their viewpoints, and consequently, isolating them in a "bubble" of their own ways of thinking and values. In addition to a filter bubble, an echo chamber is also mentioned as one of the phenomena that are said to be caused by the interaction between the intrinsic human tendency and the characteristics of Internet media. While there are risks described above, AI also has a benefit that provides personalized and filtered answers to AI business users and non-business users enabling to offer proposals in a beneficial manner.

[26] The "Expert Group on How to Ensure Healthy Information Distribution in the Digital Space" was convened by the Ministry of Internal Affairs and Communications to consider future policy directions and specific measures for ensuring the soundness of information distribution in the digital space, including responses to the distribution and spread of disinformation and misinformation on the Internet. The "Final Report" was published in September 2024.
(https://www.soumu.go.jp/menu_news/s-news/01ryutsu02_02000417.html)
In addition, as a technical response, the Ministry of Internal Affairs and Communications has been conducting the "Development and Demonstration Projects for Countermeasure Technologies against False or Misleading Information on the Internet."
(https://www.soumu.go.jp/main_sosiki/joho_tsusin/d_syohi/taisakugijutsu.html)

[27] The use of Retrieval-Augmented Generation (RAG) is expected to suppress hallucinations and improve transparency in the output process and rationale.

[28] Relevant stakeholder: The AI business actors who are directly or indirectly involved in AI use including AI developers, AI providers, AI business users, and non-business users. (The same shall apply hereafter.)

measures like model weight reduction and the selective use of models according to their purpose.)

Each AI business actor is expected to consider these matters as preconditions and to enhance the performance (usefulness) of AI as much as possible to provide people with benefits and richness and achieve well-being.

2) Safety

Each AI business actor should avoid damage to the lives, bodies, minds, and properties of stakeholders during the development, provision, and use of AI systems and services. In addition, it is important that the environment is not damaged.

(1) Taking into consideration the lives, bodies, properties and minds of humans and the environment

- ✧ Ensure that the AI system/service is sufficiently fulfilling the requirements, including the accuracy of outputs (reliability).
- ✧ Ensure that the performance level is maintained under various circumstances, and that grossly incorrect judgments are not made for irrelevant events (robustness).
- ✧ Ensure controllability that allows humans to control AI as necessary including periodic and objective monitoring and handling, in accordance with the characteristics and purposes of the relevant AI, in light of the severity and possibility of rights violations that can result from AI use or unintended AI behaviors.
- ✧ Conduct appropriate risk analyses to take countermeasures against risks (avoidance, mitigation, transference, or acceptance).
- ✧ If there are potential hazards to the lives, bodies, properties, and minds of humans and the environment, organize measures to be taken in advance and offer related information to stakeholders.
- ✧
  - Clearly specify measures that should be taken by relevant stakeholders and the terms of use.
- ✧ Determine the responses for cases where the safety of AI systems or services is endangered so that the steps can be quickly taken in such cases.

(2) Proper use (of AI)

- ✧ Develop, provide, or use AI systems and services within the range in which the AI business actor can control, preventing damage due to a provision or use that deviates from the intended purpose[29].
- ✧ For the outputs of AI systems and services centered around multimodal generative AI, it is relatively easy to generate more sophisticated outputs. Therefore, be aware that this sophistication may contribute to misunderstandings or biases and could potentially infringe on others' intellectual property rights, and implement measures[30] such as human intervention in the use of these systems[31, 32].

[29] There is a high possibility that using Retrieval-Augmented Generation (RAG) accelerates the convergence of responses generated by AI. Therefore, it is important to note that, for tasks that require content diversity and originality, the use of RAG may not be appropriate.

[30] As a measure to involve human judgment, it is suggested to include watermarks indicating that the content is AI-generated and to improve the user interface.

[31] When utilizing AI to generate program code, it is important to prioritize the safety and security of the generated code.

[32] As for the relationship with the laws with regard to intellectual properties, discussions are in progress in the Cabinet Office and the Agency for Cultural Affairs, so pay close attention to the consideration status in the future. As for the relationship between AI and copyrights, especially, the Legal System Subcommittee of the Copyright Subdivision of the Culture Council is arranging their discussions, so it is important that each AI business actor to consider response policies based on the intent of these discussions.

・Agency for Cultural Affairs, "On the Perspective of AI and Copyright" (Subcommittee on Legal Systems of the Copyright

(3) Proper training[33]

- ✧ In accordance with the characteristics and purposes of AI systems and services, ensure the accuracy, and recency as necessary, of the data (appropriateness of the data) to be used for training.
- ✧ Properly take actions such as the securement of transparency of data used for training, compliance with the legal framework, and update of AI models, within reasonable extent.
- ✧ Pay close attention to ensuring that illegal content, such as infringing reproductions and information[34], is not included in training data[33].

3) Fairness

During the development, provision, or use of an AI system or service, it is important that each AI business actor makes efforts to eliminate unfair and harmful bias and discrimination against any specific individuals or groups based on race, gender, national origin, age, political opinion, religion, and so forth. It is also important that before developing, providing, or using an AI system or service, each AI business actor recognizes that there are some unavoidable biases even if such attention is paid, and determines whether the unavoidable biases are allowable from the viewpoints of respect for human rights and diverse cultures.[35]

(1) Consideration for bias in technologies forming AI models

- ✧ There are a broad range of factors that can produce an bias, so identify the factors that might produce biases that can be considered as problems from the viewpoint of fairness. Those factors may include technological elements (algorithm, training data, AI model training process, prompts entered by AI business users or non-

---

Subcommittee, Council for Cultural Affairs, March 2024)
https://www.bunka.go.jp/seisaku/bunkashingikai/chosakuken/pdf/94037901_01.pdf

・Cabinet Office, "Interim Report of the Study Group on Intellectual Property Rights in the AI Era" (Intellectual Property Strategy Promotion Bureau, May 2024)
https://www.kantei.go.jp/jp/singi/titeki2/chitekizaisan2024/0528_ai.pdf

・Agency for Cultural Affairs, "Checklist & Guidance on AI and Copyright" (Copyright Division, Agency for Cultural Affairs, July 2024)
https://www.bunka.go.jp/seisaku/chosakuken/pdf/94097701_01.pdf

・Cabinet Office, "Interim Report of the Study Group on Intellectual Property Rights in the AI Era - Guide (for Rights Holders)" (Intellectual Property Strategy Promotion Bureau, November 2024)
https://www.kantei.go.jp/jp/singi/titeki2/chitekizaisan2024/2411_tebiki.pdf
The "Interim Report of the Study Group on Intellectual Property Rights in the AI Era" and the "Checklist & Guidance on AI and Copyright" organize examples of expected initiatives for each entity, including "AI developers," "AI providers," and "AI users,"
as well as "non-business users (general users)" and "rights holders," which differ from these guidelines.

[33] It is important that AI providers and AI business users, in addition to AI developers, also make efforts to ensure safety if they make adjustments or conduct re-trainings.

[34] In "General Understanding on AI and Copyright in Japan" (The Legal Subcommittee under the Copyright Subdivision of the Cultural Council, Agency for Cultural Affairs, March 2024)
(https://www.bunka.go.jp/seisaku/bunkashingikai/chosakuken/pdf/94037901_01.pdf), it is noted that one of the concerns raised by rights holders is that illegally uploaded content, including pirated copies, may also be used as training data.

[35] The term "bias" can be interpreted in various ways, as shown below, and this guideline uses it as a general term encompassing these interpretations.

・Statistical term (sampling bias, skewness, deviation, etc.)

・Psychological term (cognitive bias, which stems from assumptions and includes social biases due to societal norms among groups, emotional bias stemming from human emotions and conveniences, etc.)
In addition, the NIST's "Towards a Standard for Identifying and Managing Bias in Artificial Intelligence" (SP 1270)
(https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf)describes three categories of bias in AI: Systemic (arising from existing rules, norms, and practices, etc.), Statistical and Computational (relating to statistical and computational factors), and Human (arising from cognition, perception, habits, etc.), along with typical biases associated with each category.

business users[36], and reference information and collaborating external services used by AI models for inference) and behaviors of AI business users.

- ✧ Study the possibility that biases that stakeholders may not intend or detect might be produced depending on the characteristics and purposes of the AI system or service.

(2) Intervention by decisions made by humans

- ✧ To prevent AI from generating unfair results, consider implementing timely human interventions, rather than letting AI make the decisions alone. It is essential to implement measures to prevent human judgment from being influenced by automation bias during this process[37].
- ✧ Introduce a process for analyzing and handling the purposes, restrictions, requirements, and decisions for the AI system or service through clear and transparent methods, to see whether any biases have been produced.
- ✧ Be careful of unintended or undetectable biases and potential biases (biases not reflected in the training data)[38], and communicate with stakeholders from various backgrounds including culture or speciality for direction.

4) Privacy protection

It is important that during the development, provision, or use of an AI system or service, each AI business actor respects and protects privacy in accordance with its importance. At this time, relevant laws should be obeyed.

(1) Protection of privacy across AI systems and services in general

- ✧ Observe relevant laws, including the Act on the Protection of Personal Information[39], and formulate and announce the privacy policy of each AI business actor, to take measures to respect and protect the privacy of stakeholders, in accordance with its importance, based on the social contexts and legitimate expectations of people.
- ✧ Consider measures for privacy protection while taking into account the following matters:
  - Ensure measures based on the Act on the Protection of Personal Information.
  - Refer to international principles and standards for personal data protection.[40]

5) Ensuring security

[36] AI business users can train generative AI, including large-scale language models, for a specific task using a training method called in-context learning without updating the learned parameters in accordance with AI business users' inputs (called prompts).

[37] For necessary measures, refer to the aforementioned footnote 24.

[38] Examples of potential biases include cases where sensitive attributes such as race or gender are inferred from related information even when they are removed from the data, or where biases arise from the combination of multiple attributes (intersectional bias).

[39] Regarding the Act on the Protection of Personal Information, the Personal Information Protection Commission has been conducting a review every three years. As part of this, it is under consideration how personal consent should be when handling data solely for purposes such as statistics creation, including AI development, have been under consideration, and in January 2026, the "System Reform Policy under the Triennial Review of the Act on the Protection of Personal Information" was published. https://www.ppc.go.jp/personalinfo/3nengotominaoshi/

(January 2026 Announcement :

https://www.ppc.go.jp/files/pdf/01-1_seidokaiseihousin.pdf)

[40] AI business actors are expected to follow international guiding principles on privacy, including "OECD, Recommendation of the Council concerning Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data, OECD/LEGAL/0188" and "ISO/IEC 29100:2011 Information technology Security techniques Privacy framework." In addition, the Global Cross-Border Privacy Rules (CBPR) Forum has been established with the purpose of promoting the smooth cross-border transfer of personal data and the interoperability of regulations in countries across a broader range, and Japan joined it in April 2022 and has announced the Global CBPR Framework. As for generative AI, refer to the "Statement on Generative AI" by the G7 Data Protection and Privacy Authorities Roundtable (June 2023) and the "Resolution on Generative Artificial Intelligence Systems" by the Global Privacy Assembly (GPA) (October 2023) as well.

During the development, provision, or use of an AI system or service, it is important that each AI business actor ensures security to prevent the behaviors of AI from being unintentionally altered or stopped by unauthorized manipulations[41].

(1) Security measures relevant to AI systems and services[42]
- ✧ To maintain the confidentiality, integrity, and availability of AI systems and services and ensure safe and secure AI use constantly, take reasonable measures based on the technological level at the time.
- ✧ Understand the characteristics of AI systems and services, and examine whether the inter-system connections necessary for normal operations are properly established.
- ✧ Recognize that vulnerabilities in AI systems and services cannot be completely eliminated, given the possibility that the introduction of subtle perturbations into inference data may cause unintended decisions by relevant stakeholders.

(2) Consideration for the latest trends
- ✧ New methods for attacking AI systems and services from the outside are increasing on a daily basis. In order to address those risks, check the matters to be noted.

6) Transparency[43]

When developing, providing, or using an AI system or service, based on the social context when the AI system or service is used, it is important that each AI business actor provides stakeholders with information to the reasonable extent necessary and technically possible while ensuring the verifiability of the AI system or service.

(1) Ensuring verifiability
- ✧ In order to ensure verifiability relating to decisions made by AI, record or store logs of AI training processes, inference processes, rationales of decisions made by AI, and the like (for example, input/output generated when developing and using the AI system or service) to the extent possible based on the data amount or contents.
- ✧ Discuss method, frequency, maintenance period and so on recordings of data logs, taking into account the importance for identifying causes of accidents, and so forth, devising preventive measures, or proving requirements for responsibilities for damages, in accordance with characteristic of used technology as well as purposes.

(2) Providing relevant stakeholders with information

[41] The “[Guidelines on Technical Measures to Ensure the Security of AI] (Provisional Translation),” which aim to present examples of technical measures for ensuring the security of AI systems by their developers and providers, were formulated in March 2026 by the "AI Security Subcommittee" established under the Cyber Security Task Force of the Ministry of Internal Affairs and Communications.
https://www.soumu.go.jp/main_sosiki/kenkyu/cybersecurity_taskforce/index.html

[42] Details on methods are published in the “Guidelines for secure AI system development” (November 2023) by the National Cyber Security Centre (NCSC) of the UK as well.
https://www8.cao.go.jp/cstp/stmain/20231128ai.html

[43] Various countries define transparency in different ways. For example, the “Artificial Intelligence Risk Management Framework” by NIST (January 2023) defines it in three categories for AI: transparency (which can answer the question of “what happened” in the system), explainability (which can answer the question of “how” a decision was made in the system), and interpretability (which can answer the question of “why” a decision was made by the system and its meaning or context to the user), while the “ETHICS GUIDELINES FOR TRUSTWORTHY AI” by the European Commission (April 2019) defines it as traceability, explainability, and communication. The international standard ISO/IEC JTC1/SC42 defines the transparency as the degree to which appropriate information about the AI system is communicated to relevant stakeholders.

In addition, in the EU, the AI Act (Artificial Intelligence Act) (June 2024), transparency means that AI systems are developed and utilized in a way that ensures appropriate traceability and explainability, enabling humans to recognize when they are communicating or interacting with an AI system, adequately disclosing the AI system’s functions and limitations to users, and explaining rights to affected individuals.

In this document, matters relating to information disclosure are broadly referred to as “transparency.”

It is important to note that, in addition to definitions, the subjects of disclosure, the entities responsible for disclosure, and the purposes of disclosure can vary between countries.

- ✧ Based on the relations with AI and the nature and purpose of AI, provide and explain information summarizing the items listed below according to the knowledge and ability of each stakeholder.
  - AI systems and services in general
    - ➢ Fact that AI is used and its scope
    - ➢ Methods for data collection and annotation
    - ➢ Methods for training and evaluation
    - ➢ Information on the underlying AI models
    - ➢ Capabilities and limitations of the AI system or service, and proper/improper use by AI business users
    - ➢ Relevant laws applicable in the country/region where those provided with the AI system or service or AI business users are located
- ✧ Encourage a variety of stakeholders to engage actively through dialogues and collect various opinions on social impacts and safety.
- ✧ In addition, show actual advantages of providing or using the AI system or service and risks to relevant stakeholders.

(3) Reasonable and truthful support

- ✧ Provision of information to stakeholders described above "(2) Providing relevant stakeholders with information" doesn't necessarily assume disclosure of algorithms or source code, but it assumes providing them to the extent that satisfies social rationality based on the characteristics and uses of the technologies to be adopted while respecting privacy and trade secrets. (However, it is necessary to satisfy the demands for "(4) Improving explainability and interpretability for relevant stakeholders" and "7) Accountability" as stated later.)
- ✧ If any open technologies are used, conform to the rules specified for them.
- ✧ When disclose developed AI systems as open source, consider any potential social impacts.

(4) Improving explainability and interpretability for relevant stakeholders

- ✧ Share necessary explanations for those to be explained with actors who explain to analyze requirements of such explanation to gain relevant stakeholders' understanding and sense of safety to provide proof of AI operations.
  - AI provider: Inform the AI developer about things that are required to be explained.
  - AI business user: Inform the AI developer and AI provider about things that are required to be explained.

7) Accountability[44]

When developing, providing, or using an AI system or service, it is important that each AI business actor executes its accountability to stakeholders within reasonable extent for ensuring traceability, conforming to common guiding principles, and the like based on each AI business actor's roles and the degree of risks posed by the AI system or service.

(1) Improving traceability

- ✧ Establish a situation that allows the origin of data and decisions made during the development, provision, or use of the AI system or service to be traced forward and backward to the extent that is reasonable and technically possible.

(2) Explanation of conformity to common guiding principles

[44] While accountability is sometimes defined as explainability, the Guidelines handle information disclosure in the context of transparency, so accountability is defined as the concept of taking actual and legal responsibilities for AI and setting prerequisites for taking the responsibilities.

- ✧ Provide and explain information on how AI business actors conform to common guiding principles regularly to stakeholders, including suppliers, according to their knowledge and competence. This information summarizes, for example, the following items:
  - General
    - ➢ Whether any risk is found that prevents the common guiding principles from being implemented, to what extent it prevents the implementation of those guiding principles
    - ➢ Implementation progress of the common guiding principles
  - Human-centric
    - ➢ How disinformation is considered, and how diversity, inclusion, user support, and sustainability are ensured
  - Safety
    - ➢ Known risks relating to AI systems and services, countermeasures against them, and how to ensure safety against them
  - Fairness
    - ➢ Possibility that technological elements forming AI models will introduce bias. Those elements may include training data, AI model training process, prompts expected to be entered by AI business users or non-business users, and reference information and collaborating external services used by AI models for inference.
  - Privacy protection
    - ➢ Risks of infringements of privacy of AI business actors or stakeholders entailed by the AI system or service, countermeasures against those risks, and actions expected to be taken when the privacy breach actually occurred.
  - Ensuring security
    - ➢ Conformity to standards required to facilitate collaboration between AI systems and services or with other systems if such collaboration occurs
    - ➢ Any risks that may occur when the AI system or service collaborates with other AI systems and services via the Internet, and measures to be taken against the risks

(3) Designation of responsible persons

- ✧ Appoint someone as the person responsible for executing its accountability in each AI business actor.

(4) Sharing responsibilities among actors

- ✧ As for responsibilities shared among actors, clarify who take the responsibilities through contracts or social promises (voluntary commitments) between AI business actors including non-business users.

(5) Specific actions for stakeholders

- ✧ As necessary, establish and publicly report policies, including those created by each AI business actor on AI governance or privacy in relation to risk management or safety assurance associated with the use of AI systems and services. Those policies involve social responsibilities, including sharing visions with and giving out and providing information to society and general citizens.
- ✧ As necessary, set opportunities for accepting comments from stakeholders on incorrect AI output and the like, and conduct periodic and objective monitoring of the output.

- ✧ Set policies to handle cases that might affect the interests of stakeholders[45]. Execute those policies reliably and report the progress regularly to the stakeholders as necessary.

(6) Documentation[46]

- ✧ Document and store information on the items described above and keep them available for a prescribed period whenever and wherever required and able to be referenced in a manner appropriate for their use.

The specific activities expected to be implemented by each AI business actor in cooperation with society are organized as follows.

8) Education/literacy

Each AI business actor is expected to provide the persons engaged in AI in the AI business actor with the necessary education to gain the knowledge, literacy, and ethical views to correctly understand and use AI in a socially correct manner. Each AI business actor is also expected to provide stakeholders with education, in consideration of the characteristics of AI, including its complexity and the misinformation that it may provide, and possibilities of intentional misuse of AI[47].

(1) Ensuring AI literacy

- ✧ Take the necessary steps to ensure that the persons engaged in AI in each AI business actor acquire AI literacy of the level sufficient for the engagement.

(2) Education and reskilling

- ✧ It is assumed that the division of tasks between AI and humans will change due to the expansion of generative AI use, so discuss actively about education and reskilling to promote new ways of working[48].
- ✧ Provide educational opportunities taking into account differences in knowledge and skills among generations so that various people can acquire a deeper

[45] In particular, it is desirable to take appropriate measures within the necessary scope when receiving inquiries or requests for information disclosure from rights holders of training data or those who have suffered disadvantages due to AI evaluations, judgments, recommendations, or predictions. Moreover, if a disclosure order or similar directive is issued by a court, it is required to follow the said order and undertake the necessary procedures.

[46] Regarding "documentation," there is no problem as long as records are kept using appropriate tools so that they can be easily confirmed later, and they do not necessarily need to be recorded in the paper or specific document.

[47] The Ministry of Economy, Trade and Industry and IPA published "Digital Skill Standards" (December 2022) that organized the profile of ideal human resources needed in the DX era as a guiding principle on personal studies and recruitment and education of human resources in companies. In addition, they compiled "Concept of Human Resources and Skills Needed to Promote DX in Generative AI Era" in August 2023 and added the necessity for the familiarity with directions (prompts) and capabilities to put queries and build up and verify hypotheses to the skill standards. Furthermore, by July 2024, considering the impact of the widespread adoption of generative AI, we added notes on approaches to new technologies and included generative AI-related technologies such as "large language models, image generation models, and audio generation models" in the examples of learning items for "data utilization" and "technology." Additionally, the Ministry of Economy, Trade and Industry has published a "Guidebook for the Utilization of Generative AI in Content Production," which outlines directions for the proper use of generative AI in content creation, taking into consideration the protection of intellectual property rights and related interests (July 2024). (Japanese Only) https://www.meti.go.jp/policy/mono_info_service/contents/ai_guidebook_set.pdf

・Digital Skill Standards(Japanese Only) https://www.meti.go.jp/policy/it_policy/jinzai/skill_standard/main.html

・Study group on human resources policy in the digital age (Japanese Only)
https://www.meti.go.jp/shingikai/mono_info_service/digital_jinzai/index.html

The Ministry of Internal Affairs and Communications also published "The First Step of Generative AI - Introduction to the Basic Use and Precautions of Generative AI" for citizens (beginners) who may encounter generative AI in their future lives, introducing basic knowledge of generative AI, utilization scenes, introductory usage methods, and points to note when using generative AI.
https://www.soumu.go.jp/use_the_internet_wisely/special/generativeai/

[48]The Employment Policy Study Group of the Ministry of Health, Labour and Welfare has addressed career development and skills education in response to technological changes under the theme of "Boosting Labour Productivity through the Utilization of New Technologies."
https://www.mhlw.go.jp/stf/shingi2/0000204414_00017.html

understanding of benefits of AI and enhance the resilience against risks.

(3) Support for stakeholders
- ✧ To improve the safety of the whole AI system or AI service, provide stakeholders with education and literacy advancement as necessary.

9) Ensuring fair competition

Each AI business actor is expected to maintain the fair competitive environment surrounding AI so that new businesses and services using AI are created, the sustainable economic growth is maintained, and solutions for social challenges are provided.

10) Innovation

Each AI business actor is expected to make efforts to actively contribute to the promotion of innovation for the whole society.

(1) Promoting open innovation, etc.
- ✧ Promote internationalization, diversification, collaboration among industry, academia, and government sectors, and open innovation.
- ✧ Make efforts to maintain the environment in which data necessary for AI innovation is created.

(2) Consideration for interconnectivity and interoperability
- ✧ Ensure the interconnectivity and interoperability between your AI systems/services and other AI systems/services.
- ✧ When there are standard specifications, comply with them.

(3) Providing information appropriately
- ✧ Provide necessary information to the extent that does not hinder the innovation of the information provider.

In addition to the matters described above, important matters for AI developers, AI providers or AI business users, respectively, are organized in "Table 1. Important matters for each AI business actor in addition to common guiding principles." As for the matters expressed as "-" in the table, each AI business actor is expected to implement the actions described in the "Part 2. C. Common guiding principles" column, rather than doing nothing.

Hereinafter, each matter (item) described in "Table 1. Important matters for each AI business actor in addition to common guiding principles" will be identified and indicated with the notation [AI business actor - Guiding principle number) Description.].

- An AI business actor is indicated by its initial: AI **D**eveloper, AI **P**rovider, and AI Business **U**ser. A guiding principle number and description number are indicated by numbers, respectively, given in the table.

"D-2) i.", for example, refers to the important matter for AI developers about the proper data training regarding safety.

Table 1. Important matters for each AI business actor in addition to common guiding principles

	Part 2. C. Common guiding principles	Important matters for each AI business actor in addition to common guiding principles		
		Part 3. AI Developer (D)	Part 4. AI Provider (P)	Part 5. AI Business User (U)
1) Human-centric	(1) Human dignity and autonomy of individuals (2) Paying attention to manipulations by AI on decision-makings and emotions (3) Countermeasures against disinformation (4) Ensuring diversity/inclusion (5) Providing user support (6) Ensuring sustainability	-	-	-
2) Safety	(1) Taking into consideration the lives, bodies, properties and minds of humans and the environment (2) Proper use (of AI) (3) Proper training	i. Proper data training ii. Development that takes into consideration the lives, bodies, properties and minds of humans and the environment iii. Development contributing to proper use (of AI)	i. Actions against risks that consider the lives, bodies, properties, and minds of human and the environment ii. Provision contributing to proper use (of AI)	i. Proper use (of AI) that considers safety
3) Fairness	(1) Consideration for bias in technologies forming AI models (2) Intervention by decisions made by humans	i. Consideration for bias in data ii. Consideration for bias in algorithms, etc., of AI models	i. Consideration for bias in configurations and data of AI systems and services	i. Consideration for bias in input data or prompt
4) Privacy protection	(1) Protection of privacy across AI systems and services in general	i. Proper data training (Repeat of D-2) i.)	i. Deployment of mechanisms and measures for protecting privacy ii. Countermeasures against privacy violation	i. Countermeasures against inappropriate input of personal data and privacy violation
5) Ensuring security	(1) Security measures relevant to AI systems and services (2) Consideration for the latest trends	i. Deployment of mechanisms for security measures ii. Consideration for the latest trends	i. Deployment of mechanisms for security measures ii. Handling of vulnerabilities	i. Implementation of security measures
6) Transparency	(1) Ensuring verifiability (2) Providing relevant stakeholders with information (3) Reasonable and truthful support (4) Improving explainability and interpretability for relevant stakeholders	i. Ensuring verifiability ii. Providing relevant stakeholders with information	i. Documentation of system architectures and the like ii. Providing relevant stakeholders with information	i. Providing relevant stakeholders with information
7) Accountability	(1) Improving traceability (2) Explanation of conformity to common guiding principles (3) Designation of responsible persons (4) Sharing responsibilities among actors (5) Specific actions for stakeholders (6) Documentation	i. Explanation to AI providers of conformity to common guiding principles ii. Documentation of development-related information	i. Explanation to AI business users of conformity to common guiding principles ii. Documentation of service agreements or the like	i. Explanation to relevant stakeholders ii. Effective use of provided documents and conformity to agreements
8) Education/ literacy	(1) Ensuring AI literacy (2) Education and reskilling (3) Support for stakeholders	-	-	-
9) Ensuring fair competition	-	-	-	-
10) Innovation	(1) Promoting open innovation, etc. (2) Consideration for interconnectivity and interoperability (3) Providing information appropriately	i. Contribution to creation of opportunities for innovation	-	-

D. Hiroshima AI Process “Hiroshima Process International Guiding Principles for All AI Actors”

Business operators involved in all AI systems should act in accordance with the “Hiroshima Process International Guiding Principles for All AI Actors” of the Hiroshima AI Process[49].

1. We emphasize the responsibilities of all AI actors in promoting, as relevant and appropriate, safe, secure and trustworthy AI. We recognize that actors across the lifecycle will have different responsibilities and different needs with regard to the safety, security, and trustworthiness of AI. We encourage all AI actors to read and understand the “Hiroshima Process International Guiding Principles for Organizations Developing Advanced AI Systems (October 30, 2023)”[50] with due consideration to their capacity and their role within the lifecycle.

2. The following 11 principles of the “Hiroshima Process International Guiding Principles for Organizations Developing Advanced AI Systems” should be applied to all AI actors when and as relevant and appropriate, in appropriate forms, to cover the design, development, deployment, provision and use of advanced AI systems, recognizing that some elements are only possible to apply to organizations developing advanced AI systems.

Ⅰ. 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.

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

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

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

Ⅴ. 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.

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

Ⅶ. 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.

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

[49] For details, refer to “II. Hiroshima Process International Guiding Principles for All AI Actors and for Organizations Developing Advanced AI Systems” of “Hiroshima AI Process Comprehensive Policy Framework” in “Hiroshima AI Process G7 Digital & Tech Ministers’ Statement” adopted in the G7 Digital & Tech Ministers’ Meeting (December 2023).
https://www.soumu.go.jp/menu_news/s-news/01tsushin06_02000283.html
[50] https://www.soumu.go.jp/main_content/000912746.pdf

Ⅸ. 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.

Ⅹ. Advance the development of and, where appropriate, adoption of international technical standards.

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

3. In addition, AI actors should follow the 12th principle.

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

AI actors should seek opportunities to 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.

E. Building AI governance

In order to implement the common guiding principles across value chains with the cooperation of AI business actors and use AI safely and securely, it is important to build AI governance that manages risks posed by AI at levels acceptable to stakeholders and maximizes their benefits. In order to create Society 5.0, it is also essential to socially implement a system that merges the cyberspace and physical space in an advanced way (CPS) and build appropriate AI governance for the system. A society with CPS set as its foundation is complex and rapidly changes, and it is difficult to control the risks in such a society. Those social changes cause the goals targeted by the AI governance to continuously change. Therefore, it is important to practice agile governance instead of regular AI governance where the predefined rules or procedures remain unchanged. In agile governance, multiple stakeholders continuously and rapidly run a cycle consisting of environment and risk analysis, goal setting, system design, operation, and then evaluation in various governance systems in companies, regulations, infrastructure, markets, social codes and the like[51].

For specific studies, it is important to take into consideration the severity and probabilities of risks posed by the AI developed, provided, or used by each AI business actor and limitations on the resources of each AI business actor.

(1) Each AI business actor first conducts an environment and risk analysis for the AI system or service based on the benefits and risks the AI system or service may bring about during the overall lifecycle, social acceptance relating to the development and operation, changes in the external environment, and the maturity of AI.
(2) Then, according to the analysis results, each AI business actor determines whether to develop, provide, or use the AI system or service. If it decides to do so, consider setting AI governance goals[52] by establishing policies relating to AI governance. These AI

[51] Please refer to the appendix (Supplementary Materials), which provides detailed guidance for implementing AI governance, as well as "Behavioral Goals" as specific action items for each entity and hypothetical "Practical Examples" tailored to each entity.
[52] As AI governance goals, some AI business actors might establish an action policy (the name may vary with the AI business actor, for example, "AI policy") that consists of action items for the common guiding principles described in the Guidelines, whereas some AI business actors might establish an action policy that includes other elements in addition to action items for

governance goals should be consistent with each AI business actor's reason for existence and management goals such as philosophy and vision.

(3) After that, each AI business actor designs the AI management system to achieve the AI governance goals and operate the system. In this stage, each AI business actor establishes transparency and ensure accountability (such as fairness) towards external stakeholders about its AI governance goals and the operation status of those goals.

(4) Then, each AI business actor continuously monitors and evaluates whether the AI management system, including risk assessment, is effectively functioning, and make continuous improvements.

(5) After the operation of the AI system or service commences, each AI business actor repeatedly analyzes the environments and risks based on changes in the external environment, including those in the social system, such as regulations, and review the goals as necessary.

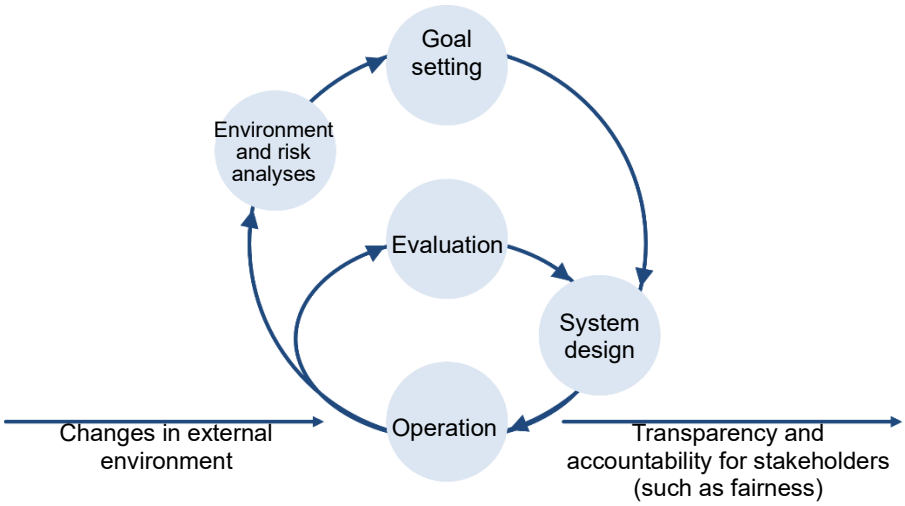

Figure 6. Basic model of agile governance

Furthermore, when studying AI governance, it is important to keep in mind the value chain and pay close attention to the following points.

- Secure the cooperation among AI business actors from the viewpoints of value chain and risk chain.
  - Example of issues among multiple AI business actors: Understanding of AI risks, improvement of quality, creation of new values through interconnections among AI systems and services (System of Systems), development of literacy of AI business users or non-business users, and so on.
  - Example of points necessary to be organized among AI business actors: Contracts concerning rights to training and using data and generated AI models.
- Clarify the risk chain including data distribution, conduct risk management activities suitable to each of the development, provision, and use stages, and build the AI governance regimes.
  - If the value and/or risk chains from AI development to service implementation are expected to span across multiple countries, understand how the international society is studying AI governance suitable for ensuring free distribution of data (Data Free Flow with Trust (DFFT)), and ensure interoperability (consisting of two

the common guiding principles (data use policy, for example). Guiding principles can also be provided to increase benefits. For example, diversity and inclusion may be improved through effective use of AI. The naming is left to each AI business actor's discretion as well.

aspects: “standard” and “interoperability between frameworks”) that is based on that study.

To make those activities effective, the management has a great responsibility, so it is important that the management exhibits leadership. It is important to think of the building of AI governance as prior investment with the aim of achieving sustainable growth and medium- and long-term expansion of each AI business actor, not to regard the building of AI governance as costs from the viewpoint of short-term pursuit of profit. Under such leadership, run the agile governance cycle shown above and fit AI governance into the strategy of each organization and the company system expecting the cycle to take hold in each organization as its culture.

Part 3 Matters Related to AI Developers

AI developers can directly design and modify AI models, so they significantly affect the output from AI as for overall AI systems and services. The society also expects them to drive innovation, so they have significant impact on the society. Therefore, it is important for AI developers to study in advance as much as possible the impacts that the AI they develop may pose when it is provided or used and take necessary measures against the impacts[53].

When developing AI, excessively focusing on accuracy may cause privacy or fairness to be compromised, or excessively focusing on privacy may cause transparency to be compromised. Thus, there may be conflict between different risks or from an ethical viewpoint. In such cases, it is important that the AI developers appropriately make decisions or corrections based on its business risks and social impacts. When an unexpected incident occurs in an AI system, any party in the AI value chain may be required to explain that incident. Bearing this in mind, it is important for AI developers to leave records that help them reasonably explain how they were involved in the AI system[54].

The matters important for AI developers are shown below.

- **During data preprocessing**
  - <u>D-2) i. Proper data training</u>
    - ✧ Properly collect training data through privacy-by-design, etc., and if it contains third-parties' personal data, data requiring attention to intellectual property rights, etc., ensure that such data is properly handled in compliance with laws and regulations throughout the lifecycle of AI ("2) Safety," "4) Privacy protection," "5) Ensuring security").
    - ✧ Implement proper protective measures before and across training by, for example, considering the deployment of any data management and restriction function that controls access to data ("2) Safety," "5) Ensuring security").

  - <u>D-3) i. Consideration for bias in data</u>
    - ✧ Take reasonable measures to control the quality of the data, noting that depending on the learning process of training data and AI models, there may be biases (including potential biases that do not appear in the training data) ("3) Fairness").
    - ✧ Based on the fact that biases cannot be completely eliminated from training data and the training process of AI models, ensure that AI models are trained with representative data sets and that AI systems are checked for the absence of unfair bias (bias that causes disadvantages to specific individuals or groups that cannot be rationally explained).

- **When developing AI**

  - <u>D-2) ii. Development that takes into consideration the lives, bodies, properties and minds of humans and the environment</u>
    - ✧ Set clear policy/guidance about safe use of AI to avoid danger incurred unexpected service/use of AI by developers ("2) Safety"):
      - Requirements for not only the performance under use conditions expected under various circumstances but also the performance achievable under the use in an unexpected environment

[53] For example, as part of efforts to ensure the trustworthiness and safety of AI, research and development is being advanced at the National Institute of Information and Communications Technology (NICT) toward the establishment of an "Active Evaluation Platform" for evaluating the trustworthiness and safety of Large Language Models (LLM).
[54] The OECD provides a catalog of tools and indicators to improve the reliability of AI development.
https://oecd.ai/en/catalogue/overview

- Requirements for methods for minimizing risks (loss of control of a linked robot, inappropriate output, etc.) (guardrail technologies, etc.)

➢ D-2) iii. Development contributing to proper use (of AI)
  ✧ Establish clear policies and guidance on how AI can be used safely in order to avoid unexpected harm caused by the provision or use of AI ("2) Safety").
  ✧ When giving a post-training to a pre-trained AI model, select a proper pre-trained AI model (whether a license for the commercial use is granted, pre-training data, specs required for the training and execution, authenticity, and so on) ("2) Safety").

➢ D-3) ii. Consideration for bias in algorithms, etc., of AI models
  ✧ Consider the possibility that bias can be included by each technical element that makes up the AI model (prompts entered by AI business users or non-business users, reference information and collaborating external services used by AI models for inference, etc.) ("3) Fairness")
  ✧ Based on the fact that biases cannot be completely eliminated from AI models, ensure that AI models are trained with representative data sets and that AI systems are checked for the absence of unfair bias ("3) Fairness").

➢ D-5) i. Deployment of mechanisms for security measures
  ✧ Throughout the development of an AI system, take security measures appropriately based on the characteristics of the adopted technologies (security by design) ("5) Ensuring security").

➢ D-6) i. Ensuring verifiability
  ✧ Note that the prediction performance and output quality of AI may significantly change or may fail to attain the expected precision after the use of AI is started. Preserve work records for follow-up verification and take measures to maintain and improve the AI quality ("2) Safety," "6) Transparency").

- **After developing AI**

➢ D-5) ii. Consideration for the latest trends
  ✧ New attack methods to AI systems are increasing on a daily basis. In order to address those risks, considerations to be noted in each step of development should be identified[55] ("5) Ensuring security").

➢ D-6) ii. Providing relevant stakeholders with information
  ✧ Provide information to relevant stakeholders in a timely manner (including cases where you provide the information via AI providers) about the AI systems that you develop ("6) Transparency"). This information may include, for example, the items listed below:
    - Possibility of changes in output or programs due to learning by AI systems ("1) Human-centric")
    - Information on safety, including technical characteristics of AI systems, mechanisms for ensuring safety, foreseeable risks that may arise as a result of using the AI system, and remedies against them ("2) Safety")
    - The expected scope of use set by AI developers in which the AI can be safely used in order to prevent harm by AI provision or use unexpected during development ("2) Safety")

[55] You can collect information via "Promotion of AI" of IPA, etc.
(Japanese Only) https://www.ipa.go.jp/digital/ai/index.html

- Information on the operational status of AI systems, causes of failures, and status of actions against them (“2) Safety”)
- Details of an update for AI, if any, and information on reasons for the update (“2) Safety”)
- Policies on collecting data learned by AI models, how AI models learn the data, and the system for implementing the learning (“3) Fairness,” “4) Privacy protection,” “5) Ensuring security”)

➢ D-7) i. Explanation to AI providers of conformity to common guiding principles
  ✧ Explain to AI providers that the prediction performance or output quality of AI may significantly change or may fail to attain the expected precision after AI starts to be used and that risks may arise as a result of this characteristic. Provide AI providers with relevant information as well. Specifically, communicate the following items (“7) Accountability”):
    - Measures against bias that technological elements forming AI models may introduce. Those elements may include training data, AI model training process, prompts assumed to be entered by AI business users or non-business users, and reference information and collaborating external services used by AI models for inference (“3) Fairness”).

➢ D-7) ii. Documentation of development-related information
  ✧ In order to improve traceability and transparency, prepare documents on your AI system development processes, data collection and labeling affecting decision-makings, algorithms you have used, and the like, as far as possible in a form that third parties can use to validate the documents (“7) Accountability”).
  (Note) This does not require to disclose all the documents prepared.

The matters at which AI developers are expected to make efforts are listed below:

➢ D-10) i. Contribution to creation of opportunities for innovation
  ✧ It is expected to implement the following items as far as possible and contribute to the creation of innovation opportunities (“10) Innovation”):
    - Research and develop quality, reliability, and development methodologies, and the like for AI.
    - Contribute to the maintenance of the sustainable economic growth and the provision of solutions for social challenges.
    - Promote internationalization, diversification, and collaboration among industry, academia, and government sectors, including watching trends in international arguments, such as DFFT, and joining AI developer communities and academic societies.
    - Provide all of society with information about AI.

Hiroshima AI Process "Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems"

AI developers developing advanced AI systems are encouraged to refer to the "Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems[56]."

I. 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.
This includes employing diverse internal and independent external testing measures, through a combination of methods for evaluations, such as red-teaming, and implementing appropriate mitigation to address identified risks and vulnerabilities. Testing and mitigation measures, should, for example, seek to ensure the trustworthiness, safety and security of systems throughout their entire lifecycle so that they do not pose unreasonable risks. In support of such testing, developers should seek to enable traceability, in relation to datasets, processes, and decisions made during system development. These measures should be documented and supported by regularly updated technical documentation.

This testing should take place in secure environments and be performed at several checkpoints throughout the AI lifecycle in particular before deployment and placement on the market to identify risks and vulnerabilities, and to inform action to address the identified AI risks to security, safety and societal and other risks, whether accidental or intentional. In designing and implementing testing 3 measures, organizations commit to devote attention to the following risks as appropriate:

- Chemical, biological, radiological, and nuclear risks, such as the ways in which advanced AI systems can lower barriers to entry, including for non-state actors, for weapons development, design acquisition, or use.
- Offensive cyber capabilities, such as the ways in which systems can enable vulnerability discovery, exploitation, or operational use, bearing in mind that such capabilities could also have useful defensive applications and might be appropriate to include in a system.
- Risks to health and/or Safety, including the effects of system interaction and tool use, including for example the capacity to control physical systems and interfere with critical infrastructure.
- Risks from models of making copies of themselves or "self-replicating" or training other models.
- Societal risks, as well as risks to individuals and communities such as the ways in which advanced AI systems or models can give rise to harmful bias and discrimination or lead to violation of applicable legal frameworks, including on privacy and data protection.
- Threats to democratic values and human rights, including the facilitation of disinformation or harming privacy.
- Risk that a particular event could lead to a chain reaction with considerable negative effects that could affect up to an entire city, an entire domain activity or an entire community.

Organizations commit to work in collaboration with relevant actors across sectors, to assess and adopt mitigation measures to address these risks, in particular systemic risks. Organizations making these commitments should also endeavor to advance research and investment on the security, safety, bias and disinformation, fairness, explainability and

[56] The "Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems" (October 2023) which was endorsed by the G7 Leaders' Statement on the Hiroshima AI Process.
Note that this document is a living document compiled based on the existing OECD AI Principles in accordance with the trends in advanced AI systems.
https://www.mofa.go.jp/mofaj/files/100573472.pdf

interpretability, and transparency of advanced AI systems and on increasing robustness and trustworthiness of advanced AI systems against misuse.

II. Identify and mitigate vulnerabilities, and, where appropriate, incidents and patterns of misuse, after deployment including placement on the market.
Organizations should use, as and when appropriate commensurate to the level of risk, AI systems as intended and monitor for vulnerabilities, incidents, emerging risks and misuse after deployment, 4 and take appropriate action to address these. Organizations are encouraged to consider, for example, facilitating third-party and user discovery and reporting of issues and vulnerabilities after deployment such as through bounty systems, contests, or prizes to incentivize the responsible disclosure of weaknesses. Organizations are further encouraged to maintain appropriate documentation of reported incidents and to mitigate the identified risks and vulnerabilities, in collaboration with other stakeholders. Mechanisms to report vulnerabilities, where appropriate, should be accessible to a diverse set of stakeholders.

III. Publicly report advanced AI systems' capabilities, limitations and domains of appropriate and inappropriate use, to support ensuring sufficient transparency, thereby contributing to increase accountability.
This should include publishing transparency reports containing meaningful information for all new significant releases of advanced AI systems.
These reports, instruction for use and relevant technical documentation, as appropriate as, should be kept up-to-date and should include, for example;
- Details of the evaluations conducted for potential safety, security, and societal risks, as well as risks to human rights,
- Capacities of a model/system and significant limitations in performance that have implications for the domains of appropriate use,
- Discussion and assessment of the model's or system's effects and risks to safety and society such as harmful bias, discrimination, threats to protection of privacy or personal data, and effects on fairness, and
- The results of red-teaming conducted to evaluate the model's/system's fitness for moving beyond the development stage.

Organizations should make the information in the transparency reports sufficiently clear and understandable to enable deployers and users as appropriate and relevant to interpret the model/system's output and to enable users to use it appropriately; and that transparency reporting should be supported and informed by robust documentation processes such as technical documentation and instructions for use.

IV. Work towards responsible information sharing and reporting of incidents among organizations developing advanced AI systems including with industry, governments, civil society, and academia
This includes responsibly sharing information, as appropriate, including, but not limited to evaluation reports, information on security and safety risks, dangerous intended or unintended capabilities, and attempts by AI actors to circumvent safeguards across the AI lifecycle.
Organizations should establish or join mechanisms to develop, advance, and adopt, where appropriate, shared standards, tools, mechanisms, and best practices for ensuring the safety, security, and trustworthiness of advanced AI systems.
This should also include ensuring appropriate and relevant documentation and transparency across the AI lifecycle in particular for advanced AI systems that cause significant risks to safety and society.
Organizations should collaborate with other organizations across the AI lifecycle to share and report relevant information to the public with a view to advancing safety, security and trustworthiness of advanced AI systems. Organizations should also collaborate and share

the aforementioned information with relevant public authorities, as appropriate. Such reporting should safeguard intellectual property rights.

V. Develop, implement and disclose AI governance and risk management policies, grounded in a risk-based approach – including privacy policies, and mitigation measures.
Organizations should put in place appropriate organizational mechanisms to develop, disclose and implement risk management and governance policies, including for example accountability and governance processes to identify, assess, prevent, and address risks, where feasible throughout the AI lifecycle.
This includes disclosing where appropriate privacy policies, including for personal data, user prompts and advanced AI system outputs. Organizations are expected to establish and disclose their AI governance policies and organizational mechanisms to implement these policies in accordance with a risk-based approach. This should include accountability and governance processes to evaluate and mitigate risks, where feasible throughout the AI lifecycle.
The risk management policies should be developed in accordance with a risk-based approach and apply a risk management framework across the AI lifecycle as appropriate and relevant, to address the range of risks associated with AI systems, and policies should also be regularly updated.
Organizations should establish policies, procedures, and training to ensure that staff are familiar with their duties and the organization's risk management practices

VI. Invest in and implement robust security controls, including physical security, cybersecurity and insider threat safeguards across the AI lifecycle.
These may include securing model weights and, algorithms, servers, and datasets, such as through operational security measures for information security and appropriate cyber/physical access controls.
This also includes performing an assessment of cybersecurity risks and implementing cybersecurity policies and adequate technical and institutional solutions to ensure that the cybersecurity of advanced AI systems is appropriate to the relevant circumstances and the risks involved. Organizations should also have in place measures to require storing and working with the model weights of advanced AI systems in an appropriately secure environment with limited access to reduce both the risk of unsanctioned release and the risk of unauthorized access. This includes a commitment to have in place a vulnerability management process and to regularly review security measures to ensure they are maintained to a high standard and remain suitable to address risks.
This further includes establishing a robust insider threat detection program consistent with protections provided for their most valuable intellectual property and trade secrets, for example, by limiting access to proprietary and unreleased model weights.

➢

VII. 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
This includes, where appropriate and technically feasible, content authentication and provenance mechanisms for content created with an organization's advanced AI system. The provenance data should include an identifier of the service or model that created the content, but need not include user information. Organizations should also endeavor to develop tools or APIs to allow users to determine if particular content was created with their advanced AI system, such as via watermarks. Organizations should collaborate and invest in research, as appropriate, to advance the state of the field.
Organizations are further encouraged to implement other mechanisms such as labeling or disclaimers to enable users, where possible and appropriate, to know when they are interacting with an AI system.

VIII. Prioritize research to mitigate societal, safety and security risks and prioritize investment in effective mitigation measures.
This includes conducting, collaborating on and investing in research that supports the advancement of AI safety, security, and trust, and addressing key risks, as well as investing in developing appropriate mitigation tools.
Organizations commit to conducting, collaborating on and investing in research that supports the advancement of AI safety, security, trustworthiness and addressing key risks, such as prioritizing research on upholding democratic values, respecting human rights, protecting children and vulnerable groups, safeguarding intellectual property rights and privacy, and avoiding harmful bias, mis- and disinformation, and information manipulation. Organizations also commit to invest in developing appropriate mitigation tools, and work to proactively manage the risks of advanced AI systems, including environmental and climate impacts, so that their benefits can be realized.
Organizations are encouraged to share research and best practices on risk mitigation.

IX. 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.
These efforts are undertaken in support of progress on the United Nations Sustainable Development Goals, and to encourage AI development for global benefit.
Organizations should prioritize responsible stewardship of trustworthy and human-centric AI and also support digital literacy initiatives that promote the education and training of the public, including students and workers, to enable them to benefit from the use of advanced AI systems, and to help individuals and communities better understand the nature, capabilities, limitations, and impact of these technologies. Organizations should work with civil society and community groups to identify priority challenges and develop innovative solutions to address the world's greatest challenges.

X. Advance the development of and, where appropriate, adoption of international technical standards
Organizations are encouraged to contribute to the development and, where appropriate, use of international technical standards and best practices, including for watermarking, and working with 8 Standards Development Organizations (SDOs), also when developing organizations' testing methodologies, content authentication and provenance mechanisms, cybersecurity policies, public reporting, and other measures. In particular, organizations also are encouraged to work to develop interoperable international technical standards and frameworks to help users distinguish content generated by AI from non-AI generated content.

XI. Implement appropriate data input measures and protections for personal data and intellectual property
Organizations are encouraged to take appropriate measures to manage data quality, including training data and data collection, to mitigate against harmful biases.
Appropriate measures could include transparency, privacy-preserving training techniques, and/or testing and fine-tuning to ensure that systems do not divulge confidential or sensitive data.
Organizations are encouraged to implement appropriate safeguards, to respect rights related to privacy and intellectual property, including copyright-protected content.
Organizations should also comply with applicable legal frameworks.

A "reporting framework" for AI developers to voluntarily verify and report their compliance with the "Code of Conduct" for developing advanced AI systems was agreed upon by the G7 in

cooperation with the OECD and has been operated since February 2025[57]. AI developers who develop advanced AI systems in compliance with the “Code of Conduct” are expected to participate in the “reporting framework.”
It should be noted that participants reported several internal benefits from engaging in the HAIP reporting process which underscore the framework’s value not only as a transparency tool but also as a mechanism for internal capacity-building and coordination[58].

- Improved coordination across teams working on trustworthy AI.
- Enabled benchmarking of governance efforts against international standards.
- Clarified internal communication on AI governance structures.
- Increased visibility into the allocation of resources across risk management areas.

[57] Hiroshima Process International Code of Conduct “Reporting Framework”
https://transparency.oecd.ai/
As of March 2026, responses have been submitted by 25 organizations, including 9 Japanese companies, and have been published on the OECD website.
https://transparency.oecd.ai/reports
[58] Ministry of Internal Affairs and Communications, Hiroshima AI Process Reporting Framework “Feedback from organisations submitting reports in the Hiroshima AI Process Reporting Framework”
https://www.soumu.go.jp/hiroshimaaiprocess/en/report.html

Part 4 Matters Related to AI Providers

AI providers are responsible for adding value to AI systems that AI developers develop and providing AI business users with AI systems and services. AI providers let AI gain popularity and expand within the society and significantly contribute to the growth of society and the economy. They have a considerable impact on society, so it is important that they provide AI systems and services on the precondition that AI is used properly. Therefore, in addition to examining whether the AI to be incorporated into an AI system or service is suited to the system or service, it is important to conduct the appropriate change management, configuration management, and service maintenance works taking into account that the expectations for AI might change in accordance with changes in the business strategy or social environment.

It is important to implement AI systems and services within the expected scope of use set by AI developers, maintain proper operation and use of the systems or services, and request AI developers to properly develop AI systems. It is important to provide AI business users with the AI service while providing and supporting the operation of the AI system or while operating the AI system. Upon provision, AI providers are expected to pay attention to prevent violations of stakeholders' rights and the occurrence of social drawbacks, etc., and share information of incidents and the like within reasonable extent to provide safer, more secure and reliable AI systems and services.

The matters important for AI providers are shown below.

- **When implementing an AI system**

  - P-2) i. Actions against risks that consider the lives, bodies, properties, and minds of human and the environment
    - ✧ Take measures that prevent AI from causing any harm on the lives, bodies, properties, and minds of stakeholders including AI business users, and the environment. The measures involve ensuring proper performances under usage conditions expected at the time of provision, enabling the AI system to maintain those performances in various situations, and minimizing (by guardrail technology or the like) risks caused by, for example, an uncontrollable robot linking to AI or improper output ("2) Safety").

  - P-2) ii. Provision contributing to proper use (of AI)
    - ✧ Establish correct considerations to note for using AI systems and services ("2) Safety").
    - ✧ Use AI within the expected scope of use set by AI developers ("2) Safety").
    - ✧ Guarantee the accuracy of AI systems/services and recency as necessary (appropriateness of data) of training data at the time of its provision ("2) Safety").
    - ✧ Examine how AI usage environments of the users of the AI system or service differ from those that AI developers expect ("2) Safety").

  - P-3) i. Consideration for bias in configurations and data of AI systems and services
    - ✧ Guarantee fairness of data at the time of its provision and examine bias contained in referenced information and collaborating external services ("3) Fairness").
    - ✧ Regularly evaluate inputs/outputs of AI models and rationales of decisions made by AI models to monitor for any bias generated. As necessary, encourage AI developers to re-evaluate the bias generated by each technical element forming AI models and promote the improvement of AI models based on the re-evaluation results ("3) Fairness").
    - ✧ Examine the possibility where bias may be introduced that arbitrarily restricts business processes and decisions made by AI business users, or non-business users on AI systems, services, or user interfaces receiving AI output results ("3)

Fairness").

- P-4) i. Deployment of mechanisms and measures for protecting privacy
  - ✧ Throughout the implementation of an AI system, take privacy protection measures by, for example, introducing a mechanism that appropriately manages and restricts access to personal data based on the characteristics of the adopted technologies (privacy by design) ("4) Privacy protection").

- P-5) i. Deployment of mechanisms for security measures
  - ✧ Throughout the provision of an AI system or AI service, take security measures appropriately based on the characteristics of the adopted technologies (security by design) ("5) Ensuring security").

- P-6) i. Documentation of system architectures and the like
  - ✧ In order to improve traceability and transparency, prepare documents describing the system architecture and data processing of the provided AI system or service that influences the decision-making ("6) Transparency").

- **After an AI system or service starts to be provided**

  - P-2) ii. Provision contributing to proper use (of AI)
    - ✧ Periodically verify whether the AI system or service is used for proper purposes ("2) Safety").

  - P-4) ii. Countermeasures against privacy violation
    - ✧ Properly collect necessary information concerning privacy protections on AI systems/services and discuss protection strategy when its violation is recognized to avoid repeated occurrence. ("4) Privacy protection").

  - P-5) ii. Handling of vulnerabilities
    - ✧ There are many new attack methods targeting AI systems and services, so identify trends in the latest risks and matters requiring attention in each provision step. And, discuss to deal with vulnerabilities ("5) Ensuring security").

  - P-6) ii. Providing relevant stakeholders with information
    - ✧ Provide information on the AI system or service to be provided (for example, the items listed below) in a timely and appropriate manner so that it can be easily understood and accessed ("6) Transparency").
      - Fact that AI is used, appropriate/inappropriate use methods, etc. ("6) Transparency").
      - Information on safety, including technical characteristics of the AI systems and services provided, foreseeable risks that may arise as a result of using the AI systems and services, and remedies against them ("2) Safety").
      - Possibility of changes in output or programs due to learning by the AI systems and services ("1) Human-centric").
      - Information on the operational status of the AI systems and services, causes of failures, status of actions against them, incidents, etc. ("2) Safety").
      - Details of an update of the AI system, if any, and information on reasons for the update ("2) Safety").
      - Policies on collecting data learned by AI models, how AI models learn the data, and the system for implementing the learning ("3) Fairness," "4) Privacy protection," "5) Ensuring security").

  - P-7) i. Explanation to AI business users of conformity to common guiding principles
    - ✧ Encourage AI business users to use AI properly and provide them with the following information ("7) Accountability"):

- Call attention to the use of data for which accuracy, and recency as necessary (appropriateness of data), are guaranteed (“2) Safety”).
- Call attention to the learning of inappropriate AI models during in-context learning (“2) Safety”).
- Precautions for when inputting personal data (“4) Privacy protection”).

✧ Call attention to inappropriate input of personal data into the AI systems and services to be provided (“4) Privacy protection”).

P-7) ii. Documentation of service agreements or the like

✧ Compile service agreements for AI business users or non-business users (“7) Accountability”).
✧ Present privacy policies (“7) Accountability”).

Incidentally, all AI actors should act in accordance with “D. Hiroshima AI Process 'Hiroshima Process International Guiding Principles for All AI Actors'” in Part 2.

Part 5 Matters Related to AI Business Users

AI providers provide AI business users with safe, secure, and reliable AI systems and services. It is important that AI business users always use the AI systems and services properly within the scope of use set by the AI providers and, as necessary, operate the AI systems. By doing so, AI business users can derive the maximum benefits from the innovation enabled by AI, including greater business effectiveness, productivity, and creativity. In addition, human intervention allows human dignity and autonomy to be conserved, helping to prevent unexpected incidents.

If AI business users are requested to explain the abilities or output results of AI by the society or stakeholders, do so to gain their acceptance by obtaining the support of AI providers. It is also expected to learn the necessary insights to use AI more effectively.

The matters important for AI business users are shown below.

- **When using AI systems and services**

  - U-2) i. Proper use (of AI) that considers safety
    - ✧ Conform to instructions for use specified by AI providers, and use AI systems and services within the expected scope of use set by AI providers during the design process (“2) Safety”).
    - ✧ Confirm whether AI systems and services are operating appropriately based on their intended specifications (“2) Safety”).
    - ✧ Input data for which accuracy, and recency as necessary (appropriateness of data), are guaranteed (“2) Safety”).
    - ✧ Understand the degrees of precision and risks of AI output and use AI output after confirming various risk factors (“2) Safety”).

  - U-3) i. Consideration for bias in input data or prompt
    - ✧ Input data for which fairness is guaranteed to avoid significant lack of fairness, pay attention to bias in prompts, and be responsible for determining whether to use AI output results for business (“3) Fairness”).

  - U-4) i. Countermeasures against inappropriate input of personal data and privacy violation
    - ✧ Refrain from improperly inputting personal data to AI systems and services (“4) Privacy protection”).
    - ✧ Collect information on privacy violation in AI systems and services properly and take the necessary steps to prevent violations (“4) Privacy protection”).

  - U-5) i. Implementation of security measures
    - ✧ Conform to instructions for security specified by AI providers (“5) Ensuring security”).
    - ✧ Pay attention not to improperly input secured information into AI systems/services (“5) Ensuring security”).

  - U-6) i. Providing relevant stakeholders with information
    - ✧ Input data for which fairness is guaranteed to avoid significant lack of fairness, and pay attention to bias in prompts when obtaining the output result from the AI system or service. When using the output result for business decision-making, inform the relevant stakeholders about the result (“3) Fairness,” “6) Transparency”).

  - U-7) i. Explanation to relevant stakeholders

- ✧ Provide information, including instructions for proper use, in a plain and accessible manner to the reasonable extent according to the nature of the relevant stakeholders (“7) Accountability”).
- ✧ If planning to use data provided by relevant stakeholders, let the stakeholders know in advance how to provide the data and its formats based on the characteristics and use purposes of AI, contact points with the relevant stakeholders as data providers, privacy policies, and the like (“7) Accountability”).
- ✧ If intending to use the AI output result as a reference for an evaluation of a specific individual or group, notify the specific individual or group to be evaluated about the use of AI, follow procedures for guaranteeing the accuracy, fairness, transparency, etc., of the output result as recommended by the Guidelines, and make a reasonable judgment by humans taking into account automation bias. If the individual or group evaluated demands you to give an explanation, fulfill your accountability by accepting the demand (“1) Human-centric,” “6) Transparency,” “7) Accountability”).
- ✧ In accordance with the characteristics of the AI systems and services to be used, set up a help desk, at the reasonable level, that handles inquiries from relevant stakeholders to give explanations and receive requests in cooperation with the AI providers (“7) Accountability”).

➢ U-7) ii. Effective use of provided documents and conformity to agreements

- ✧ Properly store and use the documents about the AI systems and services provided by the AI providers (“7) Accountability”).
- ✧ Conform to the service agreements specified by the AI providers (“7) Accountability”).

- Incidentally, all AI actors should act in accordance with “D. Hiroshima AI Process 'Hiroshima Process International Guiding Principles for All AI Actors'” in Part 2.