International rule comparison on AI transparency

July 2024

AI Law Study Group, International Exchange Subcommittee

This report was written in Japanese and machine translated. The text of the law, originally in English, was also translated into Japanese and then machine-translated. This is why the wording of the law may differ from the original. In addition, the results of machine translation have not been reviewed, so there may be errors in translation.

Table of Contents

1. Introduction.

The International Exchange Subcommittee of the AI Law Study Group compares the rules of various countries regarding the risks of AI and analyzes the characteristics of each country's rules and the social, economic, and cultural conditions in each country that underlie them. This report deals only with the results, especially in terms of transparency.

For the purposes of this report, transparency refers to the degree to which information is provided, regardless of (1) to whom the information is provided, (2) what kind of information is provided, and (3) based on what need the information is provided. However, in situations where the guidelines of each country are introduced, the definition of transparency in the relevant guidelines may be followed.

First, we will explain the parts related to transparency from the major guidelines of each country and examine their meanings and other aspects. Finally, we will consider what should be considered when seeking transparency in AI and how transparency should be in future rulemaking.

In recent years, many countries have established rules on AI, sometimes with enforceable hard law rules, as in the EU, and sometimes with unenforceable soft law rules. Both are treated here as rules.

The rules in each country differ in content due to various factors such as culture, social norms, conditions related to the use of AI, and industry conditions in each country. While there are no major differences in the value base of transparency and the need to address bias, there are differences when examining the specifics, such as what information needs to be disclosed as part of transparency.

We are examining the existence of these differences and the reasons/backgrounds for these differences. Although our examination is inadequate in terms of the countries covered and the AI principles covered, it is of a certain value and in a fast-moving world, so we are publishing it in the form of an interim report.

2. EU

2.1 Overview of AI-related measures in the EU

The European Union (EU) has insisted on the need for AI governance from an early stage, first notifying the EU Commission (the "European Commission") to the EU Parliament (the "European Parliament"), EU Council (the "EU Council"), etc. in 2018 and 2019 that guidelines The need for the introduction of the Guidelines was notified to the EU Parliament, the EU Council, and others. Based on this, the High-Level Expert Group on AI, an organization within the European Commission, published the Ethics guidelines for trustworthy AI in 2019. Here, the transparency principle is described in detail. Subsequently, preparatory work on a first draft of the EU AI Regulation has been underway in parallel with the publication of the report and white paper in 2020.

The first draft of the EU AI Regulation, which sets transparency legislation head-on, was published in April 2021. In parallel with the legislation of the EU AI Regulation, the Product Liability Directive (PLD) will be amended, the Artificial Intelligence Liability Directive (AILD) will be amended, the Product Liability Directive (PLD) will be amended, and the AILD will be published in the Official Journal in July 2024 (expected). (PLD) (Product Liability Directive) and the Artificial Intelligence Liability Directive (AILD) (AI Liability Directive) have been being drafted in parallel with the legislation of the EU AI Regulation.

In 2024, Living guidelines on the responsible use of generative AI in research and Guidelines on generative artificial intelligence and personal data for EU institutions, bodies, offices and agencies (EUIs) were published. These guidelines also refer to transparency.

2.1.1 EU Regulations and EU Directives (including draft stage)

- EU Artificial In telligence Act (EU AI Regulation) [adopted by European Parliament, Council of the European Union, 2024].[1]
- Product Liability Directive (PLD) [adopted by European Parliament, 2024].[2]
- Proposal for an Artificial Intelligence Liability Directive (AILD) (COM(2022)

[1] https://www.consilium.europa.eu/en/press/press-releases/2024/05/21/artificial-intelligence-ai-act-council-gives-final-green-light- to-the-first-worldwide-rules-on-ai/

[2] https://www.europarl.europa.eu/news/en/press-room/20240308IPR18990/defective-products-revamped-rules-to-better-protect-consumers- from-damages

496) [EU Commission, 2022].[3]

- (Reference) General Data Protection Regulation (GDPR)

2.1.2 pact

- Framework Convention on Artificial Intelligence and Human Rights AI [Council of Europe, adopted 2024].[4]

2.1.3 notification

- Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions - Artificial Intelligence for Europe (COM(2018) 237) [EU Commission, 2018].[5]
- Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions - Building Trust in Human-Centric Artificial Intelligence (COM(2019) 168) [EU Commission, 2019].[6]

2.1.4 Reports and White Papers

- White Paper on Artificial Intelligence : a European approach to excellence and trust (COM(2020) 65) [EU Commission, 2020].[7]
- Report from the Commission to the European Parliament, the Council and the European Economic and Social Committee on the safety and liability implications of Artificial Intelligence, the Internet of Things and robotics [EU Commission, 2020].[8]

2.1.5 Guidelines and others

[3] https://commission.europa.eu/business-economy-euro/doing-business-eu/contract-rules/digital-contracts/liability-rules-artificial- intelligence_en

[4] https://www.coe.int/en/web/artificial-intelligence/the-framework-convention-on-artificial-intelligence

[5] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2018%3A237%3AFIN

[6] https://digital-strategy.ec.europa.eu/en/library/communication-building-trust-human-centric-artificial-intelligence

[7] https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en

[8] https://commission.europa.eu/publications/commission-report-safety-and-liability-implications-ai-internet-things-and-robotics-0_en

- Ethics guidelines for trustworthy AI [High-Level Expert Group on AI, 2019].[9]
- Framework of ethical aspects of artificial intelligence, robotics and related technologies [EU Parliament, 2020].[10]
- Living guidelines on the responsible use of generative AI in research [EU Commission and European Research Area Forum, 2024].[11]
- Guidelines on generative artificial intelligence and personal data for EU institutions, bodies, offices and agencies (EUIs) [European Data Protection Supervisor (EDPS) (European Data Protection Supervisory Service), 2024][12]

2.1.6 AI-related measures that refer to the transparency principle

Of the above, this paper will pick out the following three that refer to the transparency principle in detail and, in the following order, provide an overview of each and how it refers to the transparency principle. In addition, a comparison will be made with the General Data Protection Regulation (GDPR), a regulation that focuses on the protection of personal data rather than AI, but is one of the laws and regulations that emphasize transparency.

- Ethics guidelines for trustworthy AI
- EU Artificial Intelligence Act (EU AI Regulation)
- Guidelines on generative artificial intelligence and personal data for EU institutions, bodies, offices and agencies (EUIs)

2.2 Position of the transparency principle in the Ethics guidelines for trustworthy AI

2.2.1 Overview of Ethical Guidelines for Trustworthy AI

In April 2019, the AI High-Level Expert Group (HLEG) created within the European Commission published the Ethical Guidelines for Trustworthy AI (the "Ethical Guidelines"), and in July 2020, based on test feedback, a revised assessment list for said Guidelines was published. At the same time, in February 2020, the EU published the White Paper on Artificial Intelligence a European approach to excellence and trust (COM(2020) 65) (AI White Paper: European

[9] https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[10] https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2020)654179

[11] https://research-and-innovation.ec.europa.eu/news/all-research-and-innovation-news/guidelines-responsible-use-generative-ai- research-developed-european-research-area-forum-2024-03-20_en

[12] https://www.edps.europa.eu/press-publications/press-news/press-releases/2024/edps-guidelines-generative-ai-embracing-opportunities -protecting-people_en

approach to excellence and trust) and the ”European Data Strategy” were published, leading to a clear goal for Europe to ”become a world leader in safely available AI”.

The Ethics Guidela in first states that a ”trustworthy AI” has the following three elements, which must be met throughout the system’s lifecycle

① Comply with all applicable laws and regulations and be lawful (Lawful AI).

② Be ethical (Ethical AI), ensuring adherence to ethical principles and values.

③ Be robust from a technical and social perspective (Robust AI), because even with good intentions, AI systems can cause unintended harm.

The Ethical Guidelines are intended to provide guidance on fostering and ensuring elements (2) and (3) of these factors.

To realize ”trustworthy AI,” the report presents a framework consisting of five components: (1) basic human rights, (2) ethical principles, (3) requirements for trustworthy AI, (4) technical and non-technical means to achieve trustworthy AI, and (5) a list of assessments of trustworthy AI.[13]

With regard to (2) ethical principles, the Ethical Guidelines explain that they consist of four principles: ”Respect for human autonomy,” ”Prevention of harm,” ”Fairness,” and ”Explicability. The four principles are ”Respect for human autonomy,” ”Prevention of harm,” ”Fairness,” and ”Explicability.

Regarding accountability here, the Ethical Guidelines state, ”It is critical to building and maintaining user confidence in AI systems. This means that processes need to be transparent, the capabilities and objectives of AI systems need to be openly communicated, and, whenever possible, decisions need to be accountable to those directly and indirectly affected. Without such information, decisions cannot be legitimately challenged.” The importance of this is emphasized as follows.

This idea seems to be based on the same kind of philosophy as the General Data Protection Regulation (GDPR), which, in Chapter 3, mainly in Articles 12, 13, and 14, strictly defines the obligation to provide information to the data subject (principal) of personal data to ensure that decisions by the data subject (principal) are based on sufficient and accurate information, and to create the preconditions for the exercise of the right to access, object, correct, and delete his or her own Although the GDPR does not define transparency itself, it is intended to promote cooperation among data protection authorities in each EU member state and to ensure the consistent implementation of the GDPR, The

[13] Yusuke Koizumi, ”AI Ethical Guidelines and Principles of Transparency Up,” Nikkan Kogyo Shimbun, June 21, 2019.

"Guidelines on Transparency" published by the European Data Protection Board (EDPB, European Data Protection Council), which aims to ensure consistent implementation of the GDPR, explains its significance in detail.

The black-box nature of algorithms, which is often pointed out as a characteristic of AI, i.e., even AI developers do not have a clear picture of why an AI model produced a particular output or decision (what combination of input factors contributed to it), and cannot always explain it This explanatory potential can be greatly undermined by the characteristics of AI as a "model" (i.e., a model that is not a "model"). In such a situation, if accountability of AI is to be ensured to a certain degree, it will be necessary to take measures in other directions, such as traceability, auditability, and the introduction of transparent communication regarding system capabilities. The extent to which these responses are required will largely depend on the severity of the consequences and individual circumstances when AI output is erroneous or inaccurate.

2.2.2 Transparency in Ethical Guidelines for Trustworthy AI

The Ethical Guidelines list the following requirements, along with subcategories, as the seven key requirements that must be met in order for an AI to be trustworthy.[14]

[14] Yusuke Koizumi, "AI Ethical Guidelines and Principles of Transparency Up," Nikkan Kogyo Shimbun, June 21, 2019.

requirement	subcategory
1. human agency and supervision	(1) Fundamental human rights, (2) Human agency, (3) Supervision by humans
2. technical robustness and safety	(1) resilience and security from attacks, (2) preliminary plans and general safety, (3) accuracy, and (4) certainty and reproducibility.
3. privacy and data governance	(1) privacy and data protection, (2) data quality and security, and (3) access to data.
Transparency	(1) Traceability, (2) Accountability, (3) Access to data
5. diversity, non-discrimination and equity	(1) avoidance of unfair bias, (2) accessibility and universal design, and (3) stakeholder participation.
6. social and environmental welfare	(1) sustainable and environmentally friendly AI, (2) social impact, and (3) society and democracy.
7. Accountability	(1) auditability, (2) minimizing and reporting negative impacts, (3) trade-offs, and (4) remedies.

Of these, with regard to traceability, the Ethical Guidelines state that "the data sets and processes that result in decisions by AI systems, including data collection, data labeling, and algorithms used, should be documented to the highest possible standard to enable traceability and transparency, and The document should be documented, as "this applies to decisions made by AI systems as well. By taking such measures, if it is discovered after the fact that an AI decision was incorrect, an investigation into the data sets and processes that may have led to said decision will allow the reasons for it to be determined and, in turn, prevent future mistakes.

The Ethical Guidelines list the following as information that should be documented to ensure traceability

✓ Has a means been established to ensure traceability? This includes documenting how

➢ Methods used in the design and development of algorithmic systems

Rule-based AI systems: how to program and how to build models

Learning-based AI systems: how the algorithm is trained, including what input data is collected and selected and how it is trained.

➢ Methods used to test and verify algorithmic systems

Rule-based AI systems: scenarios and cases used for testing and validation

Learning-based AI systems: information on data used for testing and validation

➢ Algorithmic system results

The results of decisions made by the algorithm and other decisions that may be brought about by different cases (e.g., other subgroups of users).

Regarding explainability, the Ethical Guidelines reiterate that it is the ability to explain both the technical processes of an AI system and the human decision-making associated with it (e.g., the application areas of the system). In order to be technically accountable, human beings must be able to understand and track the decisions made by the AI system.

The problem is that, as is often the case in the AI world, there is a trade-off between explainability and the accuracy of AI decisions, as attempts to improve the explainability of the system may reduce the accuracy of AI decisions, or to improve the accuracy of AI decisions may reduce the explainability of the system. The trade-off between explainability and the accuracy of AI decisions is that the AI may reduce explainability in order to improve the accuracy of the AI decision.

The Ethical Guidelines state that "whenever an AI system has a significant impact on people's lives, it must be possible to seek an adequate explanation of the decision-making process of the AI system" and that "such explanation should be timely and compatible with the expertise of the relevant stakeholders (e.g., the public, regulators, researchers) should be compatible with the expertise of the stakeholders involved (public, regulators, researchers, etc.)"; in addition, "an explanation should be available of the extent to which the AI system is influencing and shaping the organization's decision-making processes, the design choices of the system, and the rationale for deploying it." These will ensure transparency of the business model.

The Ethical Guidelines list the following as checkpoints to ensure accountability

✓ Did you evaluate it?

・ To what extent can we understand the decisions made by AI systems and, by extension, their consequences?

・ To what extent do system decisions influence the organization's decision-making process?

・ Why was this particular system deployed in this particular area?

・ What is the business model of the system (e.g., how will it create value for the organization)?

✓ Have you ensured that all users understand an explanation as to why the system resulted in a particular outcome as a result of a particular choice?

✓Did you design the AI system with interpretability in mind from the beginning?

・ Have you tried to research and use the simplest possible and easiest to interpret model for the application in question?

・ Have you evaluated whether the training and test data can be analyzed? Can the data be changed or updated over time?

・ Have you assessed whether the model's interpretability can be verified after training and development, or whether the model's internal workflows can be accessed?

Another element supporting transparency that the Ethical Guidelines emphasize is communication: with respect to the nature of communication between AI systems and their human users, the Ethical Guidelines state that "AI systems should not misrepresent themselves as human to their users. This implies that AI systems must be identifiable as such. Furthermore, it should give preference to human interaction and offer the option of rejecting this interaction when necessary to ensure compliance with fundamental rights. In turn, the capabilities and limitations of AI systems must be communicated to AI practitioners and end users in a manner appropriate to their use cases. This includes communication about the level of accuracy of the AI system and its limitations." (emphasis added). This concept is really expressed in Article 50(1) of the EU AI Regulation (with respect to interactive AI chat systems, the obligation to design and develop such systems so as to make it clear to users that they will be interacting with the AI system) and other provisions, discussed below.

The Ethical Guidelines list the following checkpoints to ensure that communication is appropriate

Did you inform ✓ (end) users, through disclaimers or other means, that they were interacting with an AI system and not another human being Did you label the AI system as such?

Have you established a mechanism to inform (end)users of the reasons and criteria for the results of the ✓AI system?

・ Did you communicate this clearly and plainly to your intended audience?

・ Have processes been established to consider user feedback and adapt the system?

・ Have you communicated about potential or perceived risks such as bias?

・ Has the use case considered communication and transparency to other audiences, third parties, or the general public?

Have you clarified the purpose of the ✓AI system and who or what will benefit from the product or service?

・ Have product use scenarios been identified and clearly communicated to ensure they are understandable and appropriate for the intended recipients?

・ Depending on the use case, have you considered human psychology and potential limitations such as confusion, confirmation bias, risk of cognitive fatigue, etc.?

Have you clearly communicated the features, limitations, and potential drawbacks of the ✓AI system?

・ For system development: Who will implement it in the product or service?

In the case of system implementation: System implementation for whom? (End) users or consumers?

Some of these checkpoints are reflected in the EU AI Regulation, described next.

2.3 Reference to Transparency Principle in EU AI Regulation

2.3.1 Overview of EU AI Regulations

The first draft of the EU AI Regulation was published in April 2021, based on the Ethical Guidelines and the AI White Paper: A European Approach to Excellence and Trust . The AI Regulation was groundbreaking and attracted worldwide attention as it sought to comprehensively regulate AI. After the publication of the draft AI Regulation, a public consultation procedure was conducted, and more than 300 comments were received from businesses and organizations not only in the EU but also from around the world. Based on the comments received, a revised draft was published on November 29, 2021. The amendments maintain the principles of the original draft, but make important changes, including making general-purpose AI systems in national security and R&D exempt, explaining that the use of AI in insurance constitutes a high risk, and extending the prohibition on social scoring in the private sphere.

Subsequently, the Council of the European Union (Council) adopted a progress report on the draft AI Regulation in June 2022 and a General Approach on the AI Bill in December 2022.In May 2023, within the European Parliament (EU Parliament) s Committee on Internal Market and Consumer Protection (IMCO) and the Committee on Civil Liberties, Justice and Home Affairs (LIBE) jointly presented and adopted an overall amendment. This version sought to clarify definitions, harmonize with existing laws and regulations such as GDPR, etc.

The following month, June 2023, further amendments were made and adopted by a majority vote at a plenary session of the European Parliament, followed by a trilogue between the European Commission, the European Parliament and the Council of the European Union, which resulted in a political agreement in December 2023, On February 2, 2024, after weeks of negotiations, the representatives of the Member States reached an agreement, which will be adopted by the European Council in March 2024 and by the Council in May 2024, and will be published in the Official Journal in July 2024 (expected). The promulgation is scheduled to take place in July 2024 (expected) upon publication in the Official Journal. Almost all of the provisions are expected to come into effect 24 months after publication in the Official Journal. Given the length of time it will take for the AI Regulation to come into force, the Commission intends to precede this by launching the AI Pact, an agreement that will commit AI developers to voluntarily fulfill the main obligations of the AI Regulation before it comes into force.

2.3.2 Transparency in EU AI Regulations

In the EU AI Regulation, Articles 11, 13, 50, and 53 are often pointed out as articles related to transparency. The following chart compares transparency in the above GDPR with transparency in the EU AI Regulation.[15]

[15] Based on FieldFisher ”Comparison of transparency requirements under the EU AI Act and GDPR”. (https://res.cloudinary.com/fieldfisher/image/upload/v1712848074/PDFs/Comparison_of_transparency_requirements_under_the_EU_AI_Act_ and_GDPR_119400306.1_cqoilg.pdf)

	Privacy Notice (GDPR)	Statement of Accountability (GDPR)	High Risk AI Systems - Description of Use	High Risk AI Systems - Technical Information	Notification regarding specific AI systems and general purpose AI models	General Purpose AI Model - Technical and Other Information
necessity	How personal data will be used and Explain individual rights.	A document explaining why the AI system was implemented , how it was designed/trained, and the bias control measures and monitoring system.	Instructions for use, including information on system features, functions and operation, purpose, performance, and limitations.	Documentation explaining the system’s compliance with applicable requirements.	Interacting with an AI system; audio, image video, or text output was artificially generated by AI*; ”deep-fake” images, audio, or video were artificially generated or manipulated using AI*; or text was artificially generated or manipulated for the purpose of providing information to the public regarding a matter of public interest*. generated or manipulated for the purpose of providing information of public interest* to the public.	Technical documentation of the model. At a minimum, it must include the specific elements listed (AI Secretariat and/or (Provided upon request by the competent authorities in each country). AI system providers seeking to integrate the GPAI model into their AI systems. Information and documentation to be made available. A good understanding of the model’s capabilities and limitations, and and contain certain enumerated elements.
Entity responsible for execution	Administrator (Controller)	Administrator (Controller)	Provider	Provider	Provider	Provider
Information	the person	The	deployer	competen	the person himself	AI Office / Competent

provided by	himself	individual and stakeholders		t authorities of each country		authorities in each country /AI system providers
Legal or self regulatory requirement	GDPR Articles 12 (Transparency and Procedures), 13-15 (Access to Information and Personal Data), 22 (same right)	GDPR Article 22 (automated decision-making for individuals, including profiling), other AI self-regulatory documents	EU AI Regulations Article 13.	EU AI Regulations Article 11.	EU AI Regulations Article 50.	EU AI Regulations Article 53.
Public/Private	public	public	public	Closed to the public. However, may be shared with notified agencies.	public	Closed to the public. However, may be shared with authorities.

Articles 11, 13, 50, and 53 are outlined below.

Article 11 and Article 13

The EU AI Regulation sets out the requirements to be complied with by high-risk AI systems in Articles 9 to 15, which must be complied with (Article 8). Of these, Article 11 is a clause on technical documentation, which stipulates that technical documentation must be prepared prior to placing on the market or providing services, and that such technical documentation must include the items listed in Annex IV, indicating that the AI concerned fulfills the requirements set forth in Part 3, Chapter 2 of the EU AI Regulation, furthermore, that the technical document must be updated from time to time. And by specifying that such technical documentation shall be prepared in such a way that it provides the competent authorities and notified bodies with the information necessary to demonstrate that high-risk AI systems comply with the requirements set out in this Chapter and to assess the compliance of AI systems with the requirements, in a clear and comprehensive manner, so that supervisory authorities can risk AI systems, it seeks to ensure an environment in which the risks of risk AI systems can be easily identified.

Next, Article 13 stipulates that high-risk AI systems must be designed and developed in such a way as to ensure that their operation is sufficiently transparent, requiring transparency from the design and development stages. On the other hand, it is an undeniable fact that, as a practical matter, there is a certain disparity between those who develop AI and those who use it in terms of the level of understanding and knowledge of AI, as well as the degree to which they are aware of the latest information and risk information. For this reason, the EU AI Regulation mandates the preparation of manuals (instructions for use) on AI systems so that users, i.e., deployers, can interpret and appropriately use the output of AI systems. This manual is required to contain concise, complete, accurate, and clear information that is appropriate, accessible, and understandable to the user, and must include prescribed items (e.g., identity and contact information of the provider, characteristics, capabilities, and performance limitations of the high-risk AI system). For service providers, the extent to which details of items such as characteristics, capabilities, and performance limitations of AI systems need to be disclosed in detail is an important point, which is detailed in the following excerpt of the clause (Article 13, Paragraph 3).

Article 50.

Article 50 is a clause setting forth transparency obligations for providers and deployers of certain AI systems. Specific AI systems here refer to certain categories of AI systems that do not fall under high-risk AI systems, including AI systems intended to directly interact with natural persons, AI systems that generate synthetic voice, images, video, text, etc., and emotion recognition systems. While these AI systems are subject to certain transparency obligations, they are also subject to certain exemptions (when their use is permitted by law for the detection, prevention, investigation, or prosecution of criminal offenses) and limitations on the method of disclosure (when the content clearly forms part of an artistic work, the method must not prevent the display or enjoyment of the work).), and limiting the method of disclosure (if the content clearly forms part of an artistic work, the method shall not interfere with the exhibition or enjoyment of the work), in an attempt to balance the need for transparency with the need to use AI in criminal investigations.

Article 53.

Article 53 provides that it is the obligation of providers of general purpose AI models to prepare technical documents containing, at a minimum, the information specified in Annex 11, and to prepare technical documents containing the information specified in Annex XII for downstream providers, for the purpose of providing them to the AI Office and national competent authorities upon request. The following is a summary of the contents of this document. Appendix 11 and 11 provide details to include, in addition to the basic description of the general purpose AI model, the elements of the AI model and the description of its development process in the technical document. General Purpose AI Model here is a translation of General Purpose AI Model, which demonstrates significant generality, including when trained using large amounts of self-supervised learning data on a large scale, regardless of how the model is brought to market, and can be integrated into a variety of downstream systems and applications AI models that can adequately perform a wide range of well-defined tasks that can be integrated into a wide range of downstream systems and applications. By stating that this does not include AI models that are used before they are released to the market for research, development, or prototyping activities, it is apparent that the intent is to avoid over-regulation, which would inhibit research, development, etc.

The following is a selection of excerpts from the preamble and text of the EU AI Regulation that refer to transparency.

Article 11 (Technical Documents)

Clause 11.1 of this text "Technical documentation for high-risk AI systems shall be prepared and kept up to date before the system is placed on the market or put into service. The technical documentation shall be prepared in such a way as to demonstrate that the high-risk AI system complies with the requirements specified in this paragraph and to provide the national competent authorities and notified bodies with the information necessary to assess the compliance of the AI system with these requirements in a clear and comprehensive form."

Article 11.2 of the main text: "When a high-risk AI system related to a product to which the Union harmonisation legislation listed in Section A of Annex I applies is placed on the market (place on the market) or put into service (put in service) A complete set of technical documentation shall be prepared that includes all the information specified in the preceding paragraph, as well as the information required under these laws and regulations."

Article 13 (Transparency and Provision of Information to Deployers)

Article 13.1 of the main text ”High risk AI systems shall be designed and developed in such a way as to ensure that their operation is sufficiently transparent to allow deployers to interpret and appropriately use the output of the system. The appropriate type and degree of transparency shall be ensured with a view to achieving compliance with the relevant obligations of the provider and deployer set forth in Article 3.”

Article 13.2 of this text, ”High-risk AI systems must be accompanied by instructions for use that contain concise, complete, accurate and clear information that is relevant, accessible and understandable to the deployer, in an appropriate digital format or otherwise.”

Article 13.3 of the text: ”The Instructions for Use must contain at least the following information

(a) the identity and contact information of the provider and, if applicable, its authorized representative

(b) Characteristics, capabilities, and performance limitations of high-risk AI systems:

(i) Intended purpose

(ii) Known and foreseeable circumstances that may affect the level of accuracy, including metrics, robustness, and cybersecurity referred to in Article 15, and the expected level of accuracy, robustness, and cybersecurity, for which high-risk AI systems are tested, validated, and anticipated

(iii) Known or foreseeable circumstances related to the use of a high-risk AI system for its intended purpose or under circumstances of reasonably foreseeable misuse, which may lead to risks to health and safety or to fundamental rights referred to in Article 9.2.

(iv) if applicable, the technical capabilities and characteristics of the high-risk AI system that provide information relevant to explaining its outputs

(v) where appropriate, the performance with respect to specific individuals or groups for whom the system is intended to be used

(vi) Other relevant information about the training, validation, and test data sets to be used, taking into account, as appropriate, the input data specifications or the intended purpose of the high-risk AI system

(vii) where applicable, information to enable the deployer to interpret and properly use the output of the high-risk AI system.”

(c) Any changes to the high-risk AI system and its performance that have been pre-determined by the provider at the time of the initial conformity assessment.

(d) the human monitoring measures referred to in Article 14, including technical measures introduced to facilitate the interpretation of the output of high-risk AI systems by the deployer

(e) Computational and hardware resources required, the expected useful life of the high-risk AI system, and the means of maintenance and care required to ensure proper functioning of the AI system, including the frequency thereof

(f) where relevant, a description of the mechanisms included in the high-risk AI system to enable the deployer to properly collect, store, and interpret the logs in accordance with Article 12."

Article 50 (Transparency Obligations for Providers and Deployers of Certain AI Systems)

Paragraph 1: "The provider shall ensure that AI systems intended to interact directly with natural persons are designed and developed in such a manner that such natural persons are informed that they will be interacting with the AI system unless it is obvious to a reasonably well-informed, observant, and thoughtful natural person, considering the circumstances and context of use The AI system must ensure that it is designed and developed in such a way that such natural person is informed that he or she is interacting with the AI system, unless it is obvious to a reasonably knowledgeable, observant and thoughtful natural person. This obligation does not apply to AI systems authorized by law to detect, prevent, investigate or prosecute crimes under appropriate safeguards for the rights and freedoms of third parties."

Section 2: "Providers of AI systems, including general-purpose AI systems that generate synthetic audio, image, video or text content, shall ensure that the output of the AI system is presented in machine-readable form and is detectable as artificially generated or manipulated. Providers must ensure that their technical solutions are effective, interoperable, robust, and reliable to the extent technically feasible, taking into account the particularities and limitations of various types of content, the cost of implementation, and generally accepted technical conditions as may be reflected in relevant technical standards. shall ensure that they are effective, interoperable, robust, and reliable. This obligation shall not apply if the AI system performs standard aids to editing or does not materially alter the input data provided by the deployer or its semantics, or if it is authorized by law for the detection, prevention, investigation or prosecution of a crime. . shall not apply."

Paragraph 3 "Deployers of emotion recognition or biometric classification systems shall inform natural persons exposed to them of the operation of the system and, where applicable, process their personal data in accordance with

Regulations (EU) 2016/679 and (EU) 2018/1725 and Directive (EU) 2016/680. The data must be processed in accordance with Regulations (EU) 2016/679 and (EU) 2018/1725 and Directive (EU) 2016/680 where applicable. This obligation shall not apply to AI systems used for biometric classification and sentiment recognition under appropriate safeguards for the rights and freedoms of third parties and in accordance with EU law and as permitted by law for the detection, prevention or investigation of crime."

Paragraph 4: "Deployers of AI systems that generate or manipulate image, audio or video content that constitutes deep fakes shall disclose that such content has been artificially generated or manipulated. This obligation does not apply where its use is authorized by law for the detection, prevention, investigation or prosecution of a crime. If the content constitutes part of a clearly artistic, creative, satirical, fictional or similar work or program, the obligation of transparency set forth in this paragraph is limited to disclosing the existence of such generated or manipulated content in an appropriate manner that does not interfere with the display or enjoyment of such work.

Implementers of AI systems that generate or manipulate text to be published for the purpose of informing the public about matters of public interest must disclose that the text was artificially generated or manipulated. This obligation does not apply where its use is authorized by law for the detection, prevention, investigation or prosecution of a crime, or where AI-generated content has undergone a process of human review or editorial control and a natural or legal person has editorial responsibility for the publication of the content."

Paragraph 5: "The information referred to in paragraphs 1 to 4 must be provided to the natural person concerned in a clear and distinguishable manner, at the latest at the time of first interaction or exposure. The information must conform to applicable accessibility requirements."

Paragraph 6: "Paragraphs 1 through 4 are without prejudice to the requirements and obligations set forth in Chapter III or any other transparency obligations set forth in federal or national law for deployers of AI systems."

Paragraph 7: "The AI Office encourages and facilitates the development of codes of practice at EU level to promote the effective implementation of obligations regarding the detection and display of artificially generated or manipulated content. (Omitted)"

Article 53 (Obligations of General Purpose AI Model Providers)

Paragraph 1 "A provider of a general purpose AI model shall

(a) This technical document shall contain, at a minimum, the information specified in Annex XI for the purpose of providing it to the AI Office and national competent authorities upon request.

(Without prejudice to the need to comply with and protect intellectual property rights and confidential business information or trade secrets in accordance with EU and national law, the information and documentation shall be prepared, kept up to date and made available to providers of AI systems who intend to incorporate the general purpose AI model into their AI systems. Without prejudice to the need to comply with and protect intellectual property rights and confidential business information or trade secrets in accordance with EU and national laws, the information and documentation must

(i) Ensure that providers of AI systems fully understand the capabilities and limitations of general purpose AI models and comply with their obligations under these rules.

(ii) contains, at a minimum, the elements set forth in Annex XII.

(c) Establish a policy to comply with federal laws regarding copyright and related rights, specifically identifying reservations of rights expressed in accordance with Article 4, paragraph 3 of EU Directive 2019/790, including state-of-the-art technology.

(d) prepare and make available to the public a sufficiently detailed summary of the content used to train the general purpose AI model, following a template provided by the AI Office."

Annex 11: Technical Document for Article 53(1)(a) - Technical Document for Providers of General Purpose AI Models

Section1 Information to be provided by all providers of general purpose AI models

The technical documentation referred to in Article 53(1)(a) shall include at least the following information, depending on the size of the model and the risk file

1. general description of the general purpose AI model

(a) the tasks the model is intended to perform and the type and nature of AI systems with which it can be integrated

(b) Applicable and acceptable use policies

(c) Date of release and method of distribution

(d) Architecture and number of parameters

(e) Input and output modalities (text, images, etc.) and formats

(f) Licenses

2. a detailed description of the elements of the model referred to in 1. and relevant information on the development process, including the following elements

(a) Technical means (e.g., usage, infrastructure, tools, etc.) required to integrate general purpose AI models into AI systems

(b) Design specifications for the model and learning process, including learning methodologies and techniques, key design choices including their rationale and assumptions, what the model is designed to optimize, and the relevance of different parameters, if applicable

(c) Information about the data used for training, testing, and validation (if applicable). This includes the type and source of data, curation methods (cleaning, filtering, etc.), number of data points, range, key characteristics, how the data was obtained and selected, all other means of detecting unsuitable data sources, and any identifiable biases (if applicable (if applicable) will be included.

(d) Computational resources used to train the model (e.g., number of floating point operations), training time, and other details relevant to training

(e) Known or estimated energy consumption of the model.

(With respect to (e), if the energy consumption of the model is unknown, the energy consumption may be based on information about the computing resources used.

Section2 Additional Information Provided

Additional information to be provided by providers of general purpose AI models with systemic risk.

1. a detailed description of the evaluation strategy, including evaluation results, based on available public evaluation protocols and tools or other evaluation methodologies. The evaluation strategy shall include evaluation criteria, evaluation indicators, and methodologies for identifying limitations.

2. if applicable, a detailed description of the measures introduced for the purpose of conducting model calibration, including internal and/or external hostile testing (e.g., red teaming), alignment and fine tuning.

3. a detailed description of the system architecture, if applicable. 4. a description of how the software components will work together and be integrated into the overall process.

Annex Ⅻ

Article 53.1(b) transparency information - technical documentation to be provided by providers of general purpose AI models to downstream providers who integrate their models into AI systems

The information referred to in Article 53(1)(b) shall include at least the following

1. general description of the general purpose AI model

(a) the tasks the model is intended to perform and the type and nature of AI systems with which it can be integrated

(b) Applicable and acceptable use policies

(c) Date of release and method of distribution

(d) if applicable, how the model interacts or can be used to interact with hardware or software that is not part of the model itself

(e) if applicable, the version of the relevant software associated with the use of the General Purpose AI Model

(f) Architecture and number of parameters

(g) Input and output modalities (text, images, etc.) and formats

(h) Licensing of that model

2. explanation of the elements of the model and its development process

(a) Technical means (instructions for use, infrastructure, tools, etc.) necessary to integrate general purpose AI models into AI systems

(b) Input and output modalities (text, images, etc.) and formats, and their maximum sizes (e.g., length of context window)

(c) Information about the data used for learning, testing, and validation, including, if applicable, the type, source, and curation method of the data

Paragraph 26 of the Preamble "A clearly defined risk-based approach should be followed to implement proportional and effective binding rules for AI systems. That approach should tailor the type and content of rules to the intensity and scope of risks that AI systems can create. Thus, it should prohibit certain unacceptable AI practices, establish requirements for high-risk AI systems and obligations for relevant operators, and establish transparency obligations for certain AI systems."

Paragraph 27 of the preamble, "(omitted) Transparency means that AI systems are developed and used in a manner that allows for adequate traceability and accountability, while at the same time making humans aware that they are communicating or interacting with AI systems and informing the implementer of the capabilities and limitations of the AI system and the affected people, and to properly inform them of their rights. (Abbreviations.)"

Paragraph 53 of the Preamble: "(Omitted) In order to ensure traceability and transparency, providers who consider that an AI system is not high risk based on the above conditions should prepare documentation of their assessment before such system is placed on the market or put into use, and provide such documentation to the competent authorities in each country upon request Such providers should provide such documentation to the competent authorities in their respective countries upon request. Such providers should be obliged to register their AI systems in the EU database established under this Regulation. The Commission should, after consulting the Council, provide further guidance on the actual implementation of the conditions under which AI systems listed in the Annex to this Regulation are not considered to be exceptionally high risk, together with a comprehensive list of examples of use of AI systems that are considered high risk and those that are not. Guidance should be provided."

Paragraph 66 of the preamble, "High-risk AI systems should be subject to requirements for risk management, quality and adequacy of the data sets used, technical documentation and records management, transparency and provision of information to implementers, human oversight, robustness, accuracy, and cybersecurity. These requirements are necessary to effectively mitigate risks to health, safety, and fundamental rights. These requirements are not unreasonable restrictions on trade when other less trade-restrictive measures are not reasonably available."

Preamble, paragraph 67: "(Omitted) To facilitate compliance with EU data protection legislation such as Regulation (EU) 2016/679, data governance and management practices should include transparency, in the case of personal data, as to the original purpose of data collection. (Abbreviated)."

Preamble, paragraph 72: "To address concerns related to the opacity and complexity of certain AI systems and to assist implementers in meeting their obligations under this Rule, transparency should be required for high-risk AI systems prior to market launch or service. High-risk AI systems should be designed to enable implementers to understand how AI systems work, assess their capabilities, and understand their strengths and limitations. At-risk AI systems should be accompanied by appropriate information in the form of instructions for use. Such information should include the characteristics, capabilities, and performance limitations of the AI system. This information should include known and foreseeable circumstances associated with the use of high-risk AI systems, including deployer actions that may affect system behavior or performance; circumstances that may lead to risks to health, safety, or fundamental rights from the AI system; changes that have been pre-determined by the deployer and assessed for suitability Includes information on relevant human monitoring measures, including changes that have been determined in advance by the deployer and assessed for suitability, and measures to facilitate the interpretation of the AI system's output by the deployer.

Transparency, including accompanying instructions for use, should assist deployers in the use of the system and support informed decision-making by deployers. In particular, deployers should be in a better position to correctly select the system they intend to use in light of the obligations that apply to them, be educated about intended and excluded uses, and use the AI system correctly and appropriately. To enhance the readability and accessibility of the information contained in the instructions for use, examples should be included where appropriate, e.g., regarding limitations and intended and excluded uses of the AI system. Providers should ensure that all documentation, including instructions for use, contains meaningful, comprehensive, accessible and understandable information that takes into account the needs and foreseeable knowledge of the intended deployer. Instructions for use should be prepared in a language that can be readily understood by the target deployer, as determined by the Member State concerned."

Preamble, paragraph 72: "To address concerns related to the opacity and complexity of certain AI systems and to assist implementers in meeting their obligations under this Rule, transparency should be required for high-risk AI systems prior to market launch or service. High-risk AI systems should be designed to enable implementers to understand how AI systems work, assess their capabilities, and understand their strengths and limitations. At-risk AI systems should be accompanied by appropriate information in the form of instructions for use. Such information should include the characteristics, capabilities, and performance limitations of the AI system. This information should include known and foreseeable circumstances associated with the use of high-risk AI systems, including deployer actions that may affect system behavior or performance; circumstances that may lead to risks to health, safety, or fundamental rights from the AI system; changes that have been pre-determined by the deployer and assessed for suitability Includes information on relevant human monitoring measures, including changes that have been determined in advance by the deployer and assessed for suitability, and measures to facilitate the interpretation of the AI system's output by the deployer.

Transparency, including accompanying instructions for use, should assist deployers in the use of the system and support informed decision-making by deployers. In particular, deployers should be in a better position to correctly select the system they intend to use in light of the obligations that apply to them, be educated about intended and excluded uses, and use the AI system correctly and appropriately. To enhance the readability and accessibility of the information contained in the instructions for use, examples should be included where appropriate, e.g., regarding limitations and intended and excluded uses of the AI system. Providers should ensure that all documentation, including instructions for use, contains meaningful, comprehensive, accessible and understandable information that takes into account the needs and foreseeable knowledge of the intended deployer. Instructions for use should be prepared in a language that can be readily understood by the target deployer, as determined by the Member State concerned."

Preamble, Section 102: "Software and data, including models, released under a free and open source license that can be openly shared and freely accessed, used, modified, and redistributed by users. General purpose AI models released under a free and open source license should be considered highly transparent and open if parameters, including weightings, information about the model architecture, and information about how the model is used are available to the public. A license should also be considered free and open source if it permits users to run, copy, distribute, study, modify, and improve software or data containing the model, provided that the original provider of the model is credited and the same or equivalent distribution terms are respected . should be considered."

Preamble, paragraph 104: "Providers of general purpose AI models that are released under a free open source license and whose parameters, including weights, information about the model structure, and information about how the model is used, are publicly available, should qualify for an exception with respect to the transparency requirements imposed by the (1) The model is not a systemic risk model. The model should not be released to the public, unless the model is considered to pose a systemic risk.

An open source license should not be considered sufficient reason to preclude compliance with the obligations under this rule. In any event, releasing general purpose AI models under a free and open source license does not necessarily reveal substantial information about the datasets used to train and fine-tune the models and how compliance with copyright law was ensured thereby, from compliance with transparency requirements to general The exception provided to the Purpose AI Model should not relate to the obligation to produce a summary of the content used to train the model or to develop a policy to identify and comply with the European Union copyright law, in particular the reservation of rights under Article 4(3) of Directive (EU) 2019/790 of the European Parliament and of the Council. . shall not."

Preamble, Section 107: "In order to increase the transparency of the data used to pre-study and train general purpose AI models, including text and data protected by copyright law, it is appropriate for the provider of such models to prepare and make publicly available a sufficiently detailed summary of the content used to train general purpose AI models. The summary should be made available to the public. While giving due consideration to the need to protect trade secrets and confidential business information, this summary should be generally comprehensive in its scope, not technically detailed, to facilitate parties with legitimate interests, including copyright holders, to exercise and enforce their rights under EU law. For example, list the main data collections or sets used to train the model, such as large private or public databases or data archives, and provide a descriptive account of other data sources used It is appropriate for the AI Secretariat to provide a summary template, simple It is appropriate for the AI Secretariat to provide a summary template, which should be simple, effective, and allow the provider to provide the required summary in narrative form. Paragraph 131 of the Preamble "In order to facilitate the work of the Commission and Member States in the field of AI and to increase transparency for the public, providers of high-risk AI systems other than those related to products included in the scope of the relevant existing EU harmonised legislation, as well as those AI systems listed in the high-risk use cases in the Annex to this Regulation Providers who consider that the AI systems listed in the Annex to this Regulation are not high risk under the exemption should be obliged to register information about themselves and their AI systems in the EU database established and maintained by the European Commission. Before using an AI system listed in the high-risk use cases in the Annex to this Regulation, deployers of high-risk AI systems that are public authorities should register in such database and select the system to be used. Other deployers should be given the right to do so voluntarily; this section of the EU database should be free and open to the public, and the information should be readily accessible, understandable, and machine-readable. (Abbreviations.)"

Paragraph 132 of the Preamble "Certain AI systems intended to interact with natural persons or generate content may pose a particular risk of spoofing or deception, whether or not they qualify as high risk. Therefore, in certain circumstances, the use of these systems should be subject to specific transparency obligations, without prejudice to the requirements and obligations for high-risk AI systems, and to targeted exceptions that take into account the special needs of law enforcement agencies. In particular, natural persons should be notified that they are interacting with AI systems unless it is obvious to a reasonably informed, observant, and prudent natural person, taking into account the context of the situation and use. In fulfilling that obligation, the characteristics of natural persons in vulnerable groups due to age or disability should also be considered, insofar as the AI system is intended to interact with those groups. In addition, natural persons should be notified if they will be exposed to AI systems that, by processing their biometric data, can identify, infer, or assign their feelings or intentions to specific categories. Such specific categories could relate to aspects such as gender, age, hair color, eye color, tattoos, personal characteristics, ethnic origin, personal preferences, interests, etc. Such information and notices should be provided in a format that is accessible to persons with disabilities."

Paragraph 134 of the Preamble "In addition to any technological solution employed by the AI system provider, any AI system that is used to generate or manipulate image, audio or video content that bears a striking resemblance to a real person, object, place, entity or event and that appears to a person as genuine or true (deep faking) or A Deployer that manipulates must clearly and distinguishably disclose that the content is artificially created or manipulated by indicating so in the AI output and disclosing its artificial origin. Compliance with this transparency obligation is subject, inter alia, to appropriate safeguards for the rights and freedoms of third parties, where the content is part of a clearly creative, satirical, artistic, fictional or similar work or program, provided that the use of the AI system or its output is in accordance with the Charter's not be construed as indicating an interference with the rights to freedom of expression and artistic and scientific freedom guaranteed by the Charter. In such cases, the obligation of transparency with respect to deep fakes set forth in these Rules is limited to disclosing the existence of such generated or manipulated content in an appropriate manner that does not interfere with the exhibition or enjoyment of the work, including its normal use and exploitation, while maintaining the usefulness and quality of the work. In addition, unless the AI-generated content has undergone an artificial process of review or editorial control, and a natural or legal person has editorial responsibility for the publication of the content, the same disclosure obligation is assumed with respect to AI-generated or manipulated text, to the extent that it is published for the purpose of informing the public of matters of public interest It is appropriate to assume the same disclosure obligations in these cases."

Preamble, paragraph 137: "Compliance with the transparency obligations relating to AI systems covered by this Regulation should not be interpreted as an indication that the use of AI systems or their outputs is lawful under this Regulation or other EU and Member State law, nor should it undermine any other transparency obligations for AI system, nor should it undermine any other transparency obligation on the deployer of the system."

2.4 Guidelines on generative artificial intelligence and personal data for EU institutions, bodies, offices and agencies (EUIs)

2.4.1 Overview of Guidelines on generative artificial intelligence and personal data for EU institutions, bodies, offices and agencies (EUIs)

Guidelines on generative Artificial Intelligence and personal data for EU institutions, bodies, offices and agencies (”EUI”) published by the European Data Protection Supervisor (EDPS) on June 3, 2024 (”Guidelines”). Guidelines on generative artificial intelligence (generative AI) and personal data for EU institutions, bodies, offices and agencies (EUIs) published by the EU Data Protection Supervisor (EDPS) offices and agencies (EUIs)) (”Guidelines”) aim to provide practical advice and instructions to assist EUIs in complying with their data protection obligations set out in Regulation (EU) 2018/1725 when using generative AI.

Regulation (EU) 2018/1725 is positioned as an amending act to make Regulation (EC) No 45/2001 equal to the General Data Protection Regulation (EU) 2016/679 (GDPR), the content of which is similar to that of the GDPR.

The Guidelines begin by emphasizing that the EDPS is publishing these Guidelines in its role as a data protection authority and not in its new role as an AI supervisory authority under the AI Regulation, and that the Guidelines do not conflict with the AI Regulation, and then position the Guidelines The position of the Guidelines is explained as follows.

These guidelines are not intended to cover in complete detail all issues related to the processing of personal data in the exploitation of generated AI subject to analysis by data protection authorities. It is only an initial orientation and a preliminary step toward the development of more comprehensive guidance. An expanded and updated version of this Guideline is to be published within 12 months of its publication.

These guidelines are organized in a Q&A style and consist of the following 14 Qs

1. what is generative AI?

2. Can EUI use generative AI?

3. how to know if the use of a generative AI system involves the processing (process) of personal data?

4. what is the role of the DPO in the development and implementation process of a

generative AI system?

5. the EUI is about to develop and implement a generative AI system; when should a DPIA (Data Protection Impact Assessment) be conducted?

6. when is the processing (process) of personal data in the design, development, and verification of a generative AI system legal?

7. how should the principle of data minimization be observed when using a generative AI system?

8. does the generated AI system respect the data accuracy principle?

9. if the EUI uses a generated AI system, how will it inform individuals about the processing (process) of their personal data?

10. what kind of response is required for generative AI with respect to the regulation on automated decisions as referred to in Article 24 of the GDPR?

11. how can we ensure fair treatment and avoid bias when using generative AI systems?

12. what about the exercise of individual rights?

13. what about data security?

14. key information resources

Of these, Q9 is related to transparency.

2.4.2 Transparency in the Guidelines on generative artificial intelligence and personal data for EU institutions, bodies, offices and agencies (EUIs)

Regarding Q9, "If the EUI uses a Generative AI System, how will it inform individuals about the processing of their personal data?" the Guidelines state that "An adequate information and transparency policy will help to mitigate risks to individuals and ensure compliance with the requirements of the Regulation, in particular by providing detailed information on how, when and why the EUI processes personal data in the Generative AI System. This means having comprehensive information (which must be provided by the developer or supplier, as the case may be) about the processing activities that take place at various stages of development, such as the source of the data set, curation/tagging procedures, and associated processing. In particular, the EUI must obtain appropriate and relevant information on the data sets used by providers and suppliers and ensure that such information is reliable and updated on a regular basis. Certain systems (e.g., chatbots) may require specific transparency requirements, such as informing individuals that they are interacting with AI systems without human intervention. Because the right to information includes the obligation to provide individuals with meaningful information about the logic, meaning, and possible impact of profiling and automated decision-making on the individual when such decisions are made, the EUI should be able to provide information not only about the functionality of the algorithms used, but also about the processing It is important to maintain up-to-date information about the data set. This obligation should generally apply even if the decision-making procedure is not fully automated, if it involves preparatory actions based on automated processing. the EUI, when using a generative AI system that processes personal data, must provide the data subject with all information required by the Regulation The information must be provided to the data subject. The information provided to the data subject shall be updated as necessary to ensure that the data subject is adequately informed and that the personal data can be properly managed." (emphasis added).

3. United Kingdom

3.1 Overview of AI-related measures in the UK

In A pro-innovation approach to AI regulation (the so-called ”AI White Paper”) and other documents published on March 29, 2023, the UK government clarified that, at this time, it will not discipline AI through general or comprehensive legislation covering AI technologies, but will instead adopt an approach that uses guidelines and standards that are appropriate to the circumstances in which AI is used. The AI regulation approach to AI regulation (the so-called ”AI white paper”) and other documents have made it clear that, at this time, the approach is to discipline AI through the use of guidelines and standards that are appropriate to the circumstances in which AI technologies are used. Various documents have been published so far, including the following, as well as tools and resources for AI safety assessment.

3.1.1 White Papers and Reports

- A pro-innovation approach to AI regulation (DSIT. March 2023)[16]
- A pro-innovation approach to AI regulation_ government response (Department for Science, Innovation and Technology (DSIT), February 2024)[17]
- Frontier AI: capabilities and risks - discussion paper (DSIT, October 2023)[18]
- Emerging processes for frontier AI safety (DSIT, October 2023)[19]
- Implementing the UK’s AI regulatory principles: initial guidance for regulators (DSIT, February 2024)[20]

3.1.2 guidance

- Explaining decisions made with AI (Information Commissioner’s office

[16] https://assets.publishing.service.gov.uk/media/64cb71a547915a00142a91c4/a-pro-innovation-approach-to-ai-regulation-amended-web- ready.pdf

[17] https://www.gov.uk/government/consultations/ai-regulation-a-pro-innovation-approach-policy-proposals/outcome/a-pro-innovation- approach-to-ai-regulation-government-response

[18] https://assets.publishing.service.gov.uk/media/65395abae6c968000daa9b25/frontier-ai-capabilities-risks-report.pdf

[19] ”Emerging processes for frontier AI safety” (DSIT, October 2023). (https://www.gov.uk/government/publications/emerging-processes-for-frontier-ai-safety/emerging-processes-for-frontier-ai-safety)

[20]https://assets.publishing.service.gov.uk/media/65c0b6bd63a23d0013c821a0/implementing_the_uk_ai_regulatory_principles_guidance_for_ regulators.pdf

(ICO and The Alan Turing Institute, 2022)[21]

- Guidance on AI and data protection Responsible AI Guidelines (ICO, March 2023)[22]
- Introduction to AI assurance (DSIT, February 2024)[23]
- Software and artificial intelligence (AI) as a medical device (Medicine & Healthcare products Regulatory Agency, June 2024)[24]

3.1.3 standard

- Algorithmic Transparency Recording Standard (Central Digital & Data Office, November 2021)[25]

3.1.4 Tools & Resources

- Portfolio of AI assurance techniques (Centre for Data Ethics and Innovation, June 2023)[26]
- Inspect (UK AI Safety Institute, May 2024)[27]

3.2 Positioning of the Transparency Principle in White Papers and Reports

3.2.1 AI White Papers

The UK government published the AI White Paper on March 29, 2023 to set forth the government's overall policy on AI regulation. Subsequently, based on the results of a public consultation, it published A pro-innovation approach to AI regulation; government response (”AI White Paper Government Response”) on February 6, 2024, updating and detailing its policies and plans. The government response indicates that the government intends to empower existing regulators to

[21] https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/explaining-decisions-made-with- artificial-intelligence/

[22] https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/

[23] https://assets.publishing.service.gov.uk/media/65ccf508c96cf3000c6a37a1/Introduction_to_AI_Assurance.pdf

[24] https://www.gov.uk/government/publications/software-and-artificial-intelligence-ai-as-a-medical-device/software-and-artificial- intelligence-ai-as-a-medical-device

[25] https://www.gov.uk/government/publications/guidance-for-organisations-using-the-algorithmic-transparency-recording-standard/ algorithmic-transparency-recording-standard-guidance-for-public-sector-bodies

[26] https://www.gov.uk/guidance/portfolio-of-ai-assurance-techniques

[27] https://ukgovernmentbeis.github.io/inspect_ai/

address AI risks according to the needs of their respective sectors. In addition, in response to the government's call for regulators in key sectors to develop a strategic approach to AI, each regulator has published a response policy in light of the AI White Paper[28] . The AI White Paper states that AI is characterized by two key characteristics: adaptability and autonomy. Adaptability refers to the ability of a system to learn and adapt to human behavior through learning. Adaptability refers to the property whereby a system develops new reasoning capabilities through learning, inferring patterns that are not easily identified by humans, while autonomy refers to the ability to make decisions without direct human control. AI systems and AI technologies with these characteristics are embodied in products and services in a variety of forms; the AI White Paper presents the following five principles as cross-cutting principles in the regulatory framework for AI defined in this way.

(1) Safety, security and robustness
(2) Appropriate transparency and explainability
(3) Fairness
(4) Accountability and governance
(5) Contestability and redress

In particular, (2) adequate transparency and accountability, which makes the decision-making process and impact of AI systems understandable and enhances the credibility of AI, is positioned as a necessary principle for the proper implementation of the other four principles.

According to the AI White Paper, transparency above means "the communication of appropriate information about the AI system to interested parties, including information about the purpose, timing, and manner of use of the AI system"[29] , and accountability means "the extent to which interested parties can access, interpret, and understand the decision-making process of the AI system and scope"[30] . These are references to the significance in IEEE 7001 (Standard for Transparency of Autonomous Systems).

The AI White Paper suggests that transparency and accountability requirements be applied in a flexible and situation-specific manner. The phrase "appropriate" transparency and accountability must be provided suggests that these requirements are not absolute, but should be tailored to the characteristics and risks of AI systems, and that the basic principles be applied to regulators in

[28] Regulators' strategic approaches to AI," dated May 1, 2024 (https://www.gov.uk/government/publications/regulators-strategic- approaches-to-ai/regulators-strategic-approaches-to-ai)

[29] [T]he communication of appropriate information about an AI system to relevant people (for example, information on how, when, and for which purposes an AI system is being used).

[30] [T]he extent to which it is possible for relevant parties to access, interpret and understand the decision-making processes of an AI system.

proportion to the risks posed by AI within the scope of their authority It is expected that.

3.2.2 Initial guidance for regulators

To implement the five regulatory principles outlined above in the AI white paper, the UK government has developed ”Implementing the UK’s AI regulatory principles: initial guidance for Implementing the UK’s AI regulatory principles: initial guidance for regulators”. This initial guidance provides specific recommendations for regulators regarding the principles of appropriate transparency and accountability.

First, it is proposed that regulators encourage AI developers and implementers to implement appropriate measures regarding the accountability of AI systems. This includes notifying AI developers and AI implementers when end users are interacting with AI systems, and explaining as simply as possible the purpose of the AI system, how decisions are made, and how the outputs are used.

It is also recommended that regulators clarify for each party involved in the lifecycle of AI the information that each party should share. In addition, it is recommended that the role of available technical standards addressing transparency and accountability of AI be considered in order to clarify regulatory guidance. Specifically, IEEE 7001 (Standard on Transparency of Autonomous Systems), ISO/IEC TS 6254 (Information Technology - Artificial Intelligence - Objectives and approaches for machine learning models and explainability and interpretability of AI systems)[31] , ISO/IEC CD 12792 (Taxonomy of transparency for AI systems), and other standards are recommended for consideration.

3.2.3 Emerging processes for frontier AI safety

Emerging processes for frontier AI safety,” published by the UK government prior to the AI Safety Summit in November 2023, outlines the following nine key processes for ensuring and maintaining the safety of ”frontier AI,” defined as ”advanced general-purpose AI models that allow AI companies to use AI technologies, particularly those that can perform a wide range of tasks and match or surpass the capabilities of today’s most advanced models. It outlines the following nine key processes for ensuring and maintaining the safety of ”frontier AI,” defined as ”advanced general-purpose AI models” that allow AI companies to perform a wide range of tasks, especially those that rival or surpass the capabilities of today’s state-of-the-art models.

[31] As of July 2024, the status is still under development.

(1) Responsible capability scaling: a risk management framework for scaling AI system capabilities
(2) Model Evaluation and Red Teaming: How to Assess Risk in AI Systems
(3) Model reporting and information sharing: measures to increase visibility of government AI development and implementation and enable users to make choices based on their full choice
(4) Security Controls: Cyber security and other AI system security measures
(5) Vulnerability reporting structure: process by which outsiders can identify safety and security issues in AI systems
(6) AI-generated content identifier: a tool to mitigate the creation and distribution of deceptive content by AI
(7) Prioritize research on risks posed by AI: Research process to identify and address emerging risks
(8) Preventing and monitoring model abuse: measures to identify and prevent intentional abuse of AI systems
(9) Data Entry Controls and Audits: Measures to control training data that may increase hazardous capabilities or risks.

Of the above, model evaluation and red teaming, model reporting and information sharing, vulnerability reporting structures, AI-generated content identifiers, and data input controls and auditing appear to be related to the transparency principle.

(1) Model Evaluation and Red Teaming

It points out the potential for frontier AI to increase the risk of harm related to misuse, loss of control, and other social risks, and introduces methods for assessing these risks. Model evaluation is described as a quantitative and reproducible way to measure the capabilities and characteristics of an AI system, while red teaming is described as a method to explore system vulnerabilities from an adversary's perspective. The document presents four categories of practice for these assessment methods: 1) conducting model assessments for multiple sources of risk, 2) performing assessments throughout the model life cycle, 3) allowing assessments by independent external assessors, and 4) supporting scientific advances in model assessment.

(2) Model reporting and information sharing

It mentions that transparency regarding frontier AI is important for AI benefit realization and risk mitigation, and contributes to the promotion of AI utilization by sharing best practices among organizations, making informed choices for users, and improving public trust. The document presents three practice categories for model reporting and information sharing: 1) sharing information on general model-independent risk assessment, mitigation, and management processes; 2) sharing model-specific information on specific frontier AI models; and 3) sharing appropriate information to different stakeholders. The following are presented.

（3）Vulnerability Reporting Structure

It is noted that unidentified safety and security issues (vulnerabilities) may still exist after implementation of frontier AI systems. To identify and address these vulnerabilities, the importance of a vulnerability management process that allows for external reporting is emphasized. The document presents three categories of practices related to vulnerability reporting structures: (1) establishing a vulnerability management process, (2) establishing a clear, easy-to-use, and publicly available reporting process, and (3) developing a collaborative vulnerability disclosure and information sharing mechanism.

（4）AI-generated content identifier

It points out that it can be difficult to distinguish AI-generated content from human-generated content, which may pose a risk to public safety; it explains that AI identifiers can help identify AI-generated content, but practical implementation has technical challenges and is currently not sufficiently reliable It is explained that they are not sufficiently reliable in their current state. The document presents three categories of practice for identifiers of AI-generated material: 1) researching techniques that can identify AI-generated content, 2) exploring techniques for watermarking AI-generated content that are robust against various variations, and 3) utilizing AI output databases.

（5）Prevent and monitor model misuse

It points out that the intentional misuse or abuse of AI systems can pose serious risks to individuals, organizations, and society as a whole. In particular, it emphasizes the danger that the advanced capabilities of frontier AI could be used for malicious purposes, such as fraud, cyber attacks, and the spread of disinformation. The document identifies four categories of practices related to preventing and monitoring the misuse of models: 1) identifying potential misuse and abuse scenarios for AI systems and developing defenses against them, 2) implementing systems to monitor the use of AI systems and detect abnormal or suspicious activity, 3) (iii) develop and provide users with guidelines for the appropriate use of AI systems, and (iv) establish a rapid response mechanism when misuse or abuse is detected.

(6) Data entry control and auditing

It points out that the data used to train AI systems can affect system behavior, and if frontier AIs are trained with low quality or undesirable data, risk can increase and potentially dangerous capabilities can be enhanced. It explains that data control and auditing can mitigate risk by more accurately predicting the system's capabilities and removing input data that could produce dangerous capabilities. The document identifies four categories of practices related to data input control and auditing: (1) implementing responsible data collection practices prior to training data collection, (2) auditing input data before using it to train AI systems to identify data that could produce potentially hazardous capabilities, (3) taking and take appropriate risk mitigation measures accordingly; and (4) facilitate evaluation of input data by external parties and share data audit information.

3.3 Reference to transparency principles in the Guidance

3.3.1 Introduction to AI assurance

The UK government has published guidance entitled "Introduction to AI assurance," aimed primarily at those new to the concept of AI assurance, with the goal of promoting AI assurance and helping businesses and organizations build secure and reliable AI systems The guidance is intended primarily for those who are new to the concept of AI assurance. The guidance addresses the following with respect to the practice of transparency principles.

(1) Mechanisms for Ensuring Reliability of AI

The Guidance introduces various reliability assurance measures, including risk assessment, impact assessment, bias audit, compliance audit, conformance assessment, and formal verification. Each measure aims to ensure transparency in different aspects of the AI system. For example, bias audits serve to ensure the fairness of algorithms, and compliance audits serve to ensure transparency in regulatory compliance.

It is also recommended that these mechanisms be used in combination throughout the AI lifecycle to achieve comprehensive transparency.

In addition, it is recommended that the reliability of AI be ensured in accordance with international standards such as ISO/IEC 42001 (Information technology - Artificial intelligence - Management systems), ISO/IEC TR 24027 (Bias in AI systems, decision support by AI), ISO/IEC TS 12791 (Machine learning in the

Handling of Undesirable Bias in Classification and Regression Tasks)[32] , etc.) It is recommended to achieve AI reliability assurance in line with international standards such as

(2) Scope of AI Reliability Assurance

The scope of AI reliability assurance extends to training data, AI models, AI systems, and the broader operational context, and it is noted that reliability assurance is necessary at all stages of AI development and use, from data collection to final system deployment.

(3) Assurance of data, models, systems, and governance

Methods for ensuring trustworthiness in data, models, systems, and governance elements are described in detail. For data, the establishment of transparent and standardized data collection, processing, and sharing processes is recommended, including a clear data strategy and division of responsibility for data management. For models and systems, the use of evaluation techniques such as impact assessment and performance testing is suggested to ensure transparency of their functioning and results. For governance, integration of an organization-wide AI governance framework with transparency at its core is recommended, which includes establishing clear processes for risk identification, management, and mitigation.

(4) Governance Process

The establishment of governance processes centered on transparency is strongly recommended. This includes establishing standardized internal transparency and reporting processes, with a clear delineation of responsibilities and the designation of a data controller being emphasized. External transparency and reporting processes are also listed as a core governance process, with appropriate disclosure to stakeholders recommended. In addition, specific practices are provided, such as setting milestones in project design and establishing clear pathways for escalating concerns.

3.3.2 Explaining decisions made with AI

The Information Commissioner's Office (ICO), the UK's data protection authority, aims to provide organizations with practical advice for explaining to individuals the

[32] As of July 2024, the status is still under development.

processes, services, and decisions provided or supported by AI, In collaboration with The Alan Turing Institute, the ICO has published ”Explaining decisions made with AI”. This document is also referenced in the ICO’s March 15, 2023 update of the Guidance on AI and Data Protection Responsible AI Guidelines as a key document for implementing the principles of transparency. The ”Explaining decisions made with AI” contains the following specific recommendations

(1) Selection of preferred description type

It is recommended that the preferred explanation type be selected based on the organization’s domain, the specific use case, and the impact on the individual. In many cases, ”rationale” and ”responsibility” explanations will take precedence; other explanation types may become important depending on the situation. It is considered effective to document the selected explanation type and the reasons for it, and to seek input from colleagues and customers as needed.

(2) Collection and preprocessing with data accountability in mind

Awareness of the accountability of data processing is emphasized from the data collection and preprocessing stages. Appropriate documentation of data sources, collection methods, and preprocessing procedures is required.

(3) Building Interpretable AI Systems

When building AI systems, it is recommended that models be selected that are easy to understand in terms of their internal behavior and ensure a level of interpretability appropriate to the use case and its impact on the individual. In addition, the system should be flexible enough to accommodate a variety of explanation types, utilizing supplemental explanatory techniques to the ”black box” model as needed.

(4) Comprehensible description of the statistical results of the AI model

It describes how to present the statistical results of AI models in a way that is understandable to users and decision makers. It is considered important to describe the mathematical basis in everyday language so that even non-technical stakeholders can understand it.

(5) Provide appropriate training to users of AI technology

Emphasis is placed on providing appropriate training to human decision makers who utilize AI to better understand the basics of machine learning, the limitations of AI, and automated decision support technologies.

(6) Construction of explanations and selection of appropriate presentation methods

It is recommended that the construction and presentation of explanations be carefully considered and that appropriate methods be selected for each situation, including websites, applications, written materials, and face-to-face meetings. It is recommended that the method and level of information provision be customized to take contextual factors into account, and that a hierarchical approach be used to prioritize the most relevant information and ensure that detailed explanations are easily accessible when necessary.

3.4 Standards for Transparency Principles

3.4.1 Algorithmic Transparency Recording Standard

The Algorithmic Transparency Recording Standard (ATRS) sets standards for public disclosure of information about algorithmic tools used by public sector organizations and how they support decision-making. An algorithmic tool is a product, application, or device that uses complex algorithms to support or solve a specific problem, meaning not only artificial intelligence (AI) but also statistical modeling and complex algorithms in general.

The use of ATRS is recommended when the use of a public sector algorithmic tool has a significant direct or indirect public effect on the decision-making process or interacts directly with the public. In determining whether a tool has a public effect, consider whether it has a substantial impact on an individual, organization, or group; a legal, economic, or similar impact; an impact on procedural or substantive rights; or an impact on eligibility for, receipt of, or denial of a program.

If the public sector uses ATRS, it must complete an Algorithm Transparency Report that includes the following summary description and detailed explanation, and the report will be uploaded to the GOV.UK repository.

The summary description should provide a short non-technical description of the algorithmic tool and outline how the tool works, how the tool is incorporated into the decision process, the problem you are trying to solve with the tool and its solution, why you justify using the tool, etc. The following information should be provided.

The detailed description should include the owner and responsible party of the tool, detailed functions of the tool and the reasons for its use, a description of how the tool is integrated into the decision-making process and how the tool impacts the decision-making process, technical specifications and data, and an impact assessment of using the tool.

According to the AI Whitepaper Government Response, the use of ATRS is mandatory in all central government departments, with future expansion to the broader public sector planned.

3.5 Tools and resources to put the Transparency Principles into practice

3.5.1 Portfolio of AI Assurance Techniques

The Portfolio of AI Assurance Techniques is a resource for businesses and individuals involved in the design, development, implementation, or procurement of AI systems, providing guidance and resources on the use of techniques to assess and verify the reliability of AI systems.

Each application case study provides an overview of the case study, which of the five principles presented in the AI White Paper it relates to, the approach and method of evaluation and verification, benefits of utilization, technical limitations, and other information in an organized manner. The case studies can be searched by category, such as technical field, sector, and principles presented in the AI White Paper.

3.5.2 Inspect

Inspect, developed by the UK AI Safety Institute and released on May 10, 2024, is an open source platform for evaluating the safety of AI models. Inspect can be used by OpenAI, Anthropic, Google, Mistral, Hugging Face, Ollama, TogetherAI, AWS Bedrock, Azure AI, Cloudflare, and other various AI providers offering large-scale language models (LLMs) and generative AI, making it easy to compare models across different providers.

The scoring function is a key feature of Inspect, which uses three main components: 1) a data set of sample test scenarios for evaluation, 2) a solver that runs test scenarios using prompts, and 3) a scorer that analyzes the solver’s output to generate a score for AI Evaluate the safety of the model.

(1) data-set

Datasets are the basis for the evaluation and are available in common formats such as CSV. Datasets on Hugging Face or stored in Amazon S3 are also available. The dataset contains inputs to the model, expected outputs, and metadata necessary for evaluation.

(2) solver

The solver is responsible for running the AI model on each sample of the data set. This includes setting system messages, generating prompts, executing the model, and retrieving output, ranging from simple generation to complex inference chains.

(3) scorer

The scorer determines how close the model's output is to the target answer. Methods range from simple evaluations using text matching and regular expressions to advanced evaluations that use another AI model to determine answer quality. The "model scoring" feature allows for responses to open questions and evaluation of facts embedded in longer texts. In addition, Inspect allows multiple models to be used for scoring, and a majority vote can be used to determine the final rating.

Inspect is open source, allowing developers to customize existing scoring features or add entirely new evaluation methods.

3.6 Characteristics of Measures Related to the Transparency Principle in the United Kingdom

As articulated by the AI White Paper, the UK government has adopted a gradual and cautious approach to the regulation of AI. Rather than immediately introducing legally binding laws and regulations, the policy is to first implement non-legally binding principles and then evaluate their effectiveness before considering legislation for each individual area. This strategy aims to achieve effective risk management without stifling innovation.

In particular, it is notable that while recognizing that binding measures will be necessary in the future for "highly capable general-purpose AI" (models that can perform a wide range of tasks and match or surpass the capabilities of current state-of-the-art models), it has maintained a cautious position, refraining from hasty introduction of regulations. The most distinctive feature of this approach is that it refrains from introducing regulations too hastily, while recognizing that binding measures will be necessary in the future.

In addition, the AI White Paper and other policy documents have been proactively developed to provide a comprehensive regulatory approach for regulators, centered on five regulatory principles, and each regulator presents a separate regulatory framework based on this approach to ensure consistency and harmonization of regulations. This unified approach enables operators and individuals involved in AI to understand and respond appropriately to the UK's AI policy in an integrated manner.

In addition, the UK government actively utilizes international and technical standards and encourages compliance with them. This policy is intended to help globally operating companies build AI governance efficiently and effectively by emphasizing consistency and alignment with international best practices.

In addition, the UK government is committed to developing and publishing specific and practical guidance, tools and resources. These efforts will assist domestic and international stakeholders in putting the AI transparency principles into practice at a practical level.

These approaches have resulted in a flexible and effective AI regulatory framework, which is considered to be a hallmark of the UK's AI policy.

4. United States of America

This section presents and analyzes the hard law and soft law status of AI in the United States. In particular, we organize specific efforts on AI trustworthiness from a socio-technical perspective that places people and technology in the context of an organization's business and functions.

4.1 Overview of AI-related measures in the U.S.

AI-related measures in the U.S. include hard law, soft law, and even presidential decrees. In conjunction with soft law, efforts have begun on programs and test beds to evaluate and measure AI technologies, including generative AI from a socio-technical perspective.

4.1.1 Hard Law Presidential Decree

In the United States, those classified as hard law are the National Artificial Intelligence Initiative Act of 2020[33] and the Artificial Intelligence In Government Act of 2020[34] . Also, while not hard law, the Presidential Executive Order on the Safe, Secure, and Reliable Development and Use of Artificial Intelligence (AI)[35] was issued in 2023.

4.1.2 soft law

Regarding the development of AI-related guidelines and other soft law in government, there is the Blueprint for an AI Bill of Rights (White House)[36] and the Responsible AI Guideline (DoD)[37] issued in 2022.

[33] White House, "National Artificial Intelligence Initiative Act of 2020" (March 2020) (https://www.congress.gov/bill/116th-congress/house-bill/ 6216)

[34] White House, "Artificial Intelligence In Government Act of 2020" (September 2020) (https://www.congress.gov/bill/116th-congress/house-bill/2575)

[35] White House, "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence" (October 2023) (https://www. whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and- use-of-artificial-intelligence/)

[36] White House "Blueprint for an AI Bill of Rights" (October 2022) (https://www.whitehouse.gov/ostp/news-updates/2022/10/04/blueprint-for-an-ai- bill-of-rightssa-vision-for-protecting-our-civil-rights-in-the-algorithmic-age/)

[37] Responsible AI Guidelines [DoD,2022] DoD "Responsible AI Guidelines" (May 2023) (https://www.diu.mil/responsible-ai-guidelines#overview)

NIST published the AI RISK MANAGEMENT FRAMEWORK (AI RMF)[38] in 2023, and has since published companion resources (guidance documents) and launched initiatives in accordance with Executive Order. The Profile of Generative AI (NIST AI 600-1)[39] as a companion resource for generative AI, and Generative AI and Dual-Use Infrastructure as a companion resource for the Secure Software Development Framework. Secure Software Development Procedures for Generative AI and Dual-Use Infrastructure Models (NIST SP 800-218A)[40] , as a companion resource to Reducing the Risks Posed by Synthetic Content (NIST AI 100-4)[41] , as well as Adversarial Machine Learning (NIST AI 100-2e2023)[42] , and Plan for Global Engagement on AI Standards (NIST AI 100-5)[43] have been issued.

4.1.3 Evaluation Programs and Testbeds

Programs and testbeds for evaluation have also been launched, with NIST Dioptra[44] , NIST GenAI[45] , and NIST ARIA (Assessing Risks and Impacts of AI)[46] underway.

[38] NIST, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)" (January 2023) (https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id= (936225)

[39] Profile on Generative AI (NIST AI 600-1) [NIST,2024]
NIST "Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile" ((April 2024)
(https://airc.nist.gov/docs/NIST.AI.600-1.GenAI-Profile.ipd.pdf)

[40] Secure Software Development Practices for Generative AI and Dual-Use Foundation Models (NIST SP 800-218A) [NIST,2024]
NIST "Secure Software Development Practices for Generative AI and Dual Foundation Models" (April 2024)
(https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-218A.ipd.pdf)

[41] Reducing Risks Posed by Synthetic Content (NIST AI 100-4) [NIST,2024]
NIST "Reducing Risks Posed by Synthetic Content" (April 2024) (https://airc.nist.gov/docs/NIST.AI.100-4.SyntheticContent.ipd.pdf)

[42] Adversarial Machine Learning (NIST AI 100-2e2023) [NIST, 2024]
NIST "Adversarial Machine Learning A Taxonomy and Terminology of Attacks and Mitigations" (January 2024)
(https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2023.pdf)

[43] Plan for Global Engagement on AI Standards (NIST AI 100-5) [NIST,2024] ★Draft
NIST "A Plan for Global Engagement on AI Standards" (April 2024) (https://airc.nist.gov/docs/NIST.AI.100-4.SyntheticContent.ipd.pdf)

[44] NIST "Dioptra" (April 2023)
(https://pages.nist.gov/dioptra/)

[45] NIST "Evaluating Generative AI Technologies: GenAI" (April 2024)
(https://ai-challenges.nist.gov/genai)

[46] NIST, "Assessing Risks and Impacts of AI: ARIA" (April 2024)
(https://ai-challenges.nist.gov/aria https://ai-challenges.nist.gov/aria/library)

4.2 A Socio-Technical Approach

Socio-technical (Socio-technical) is a key word that appears in the AI RMF; according to NIST SP 1270[47] , "a term used to describe how humans interact with technology in a broader social context," "transparency, data sets, testing, evaluation, verification and validation (TEVV) cannot be overemphasized. Participatory design techniques, multi-stakeholder approaches, and human involvement in the loop are also important to mitigate the risks associated with AI bias. But...each has its pitfalls. What is missing from the current remedy is guidance from a broader sociotechnical perspective that links these practices to societal values." The report states.

4.2.1 Trustworthy AI systems

The AI RMF provides an AI Risk Management Framework to better manage the risks associated with AI systems. The framework is organized around four main functions: governance, mapping, measurement, and management. It also addresses the relationship of trustworthy AI systems to social technologies, summarized in seven characteristics.

(1) Safety, (2) Security and resilience, (3) Accountability and interpretability, (4) Privacy enhancement, (5) Fairness and management of harmful bias, (6) Accountability and transparency, (7) Validity and reliability

All seven of these characteristics are socio-technical system attributes that must address such diverse criteria that are valuable to stakeholders, and neglecting them may increase the probability of negative outcomes.

Therefore, scientifically supported testing (T), evaluation (E), verification (V), and validation (V) (TEVV) should be performed periodically throughout the AI's lifecycle to provide insight regarding technical, social, legal, and ethical standards and norms. This approach to increasing the trustworthiness of AI can reduce the negative risks of AI.

4.2.2 Technical Guidelines to Support AI Risk Management

As noted above, the NIST AI RMF is the basis for the supporting technical guidelines that have been issued. These can be organized in relation to the seven characteristics of trustworthy AI systems, as shown in the figure.

The Profile on Generative AI (NIST AI 600-1) builds on the AI RMF and summarizes actions for risks specific to generative AI. It includes actions that

[47] NIST Special Publication 1270, Towards a Standard for Identifying and Managing Bias in Artificial Intelligence
(https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf)

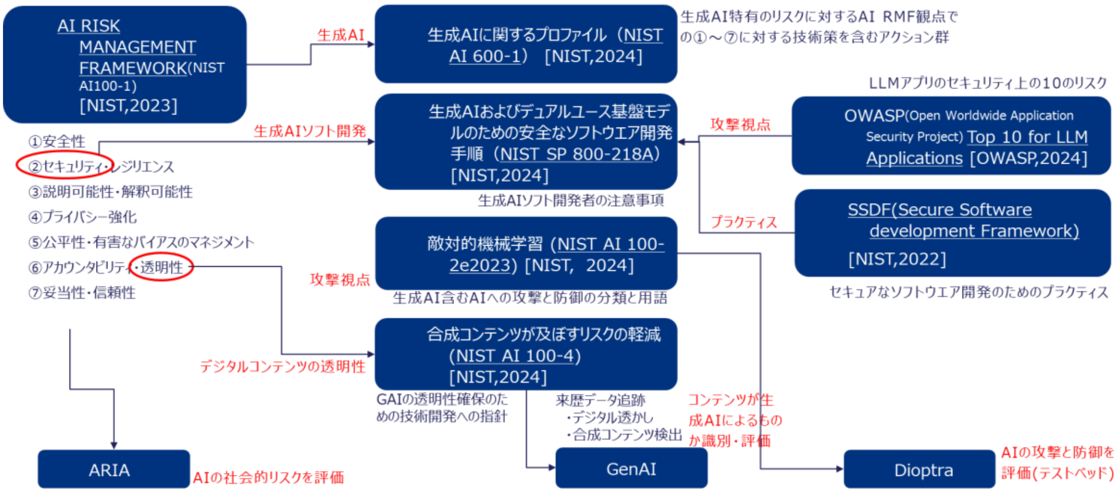
contribute to technical measures for seven characteristics of AI trustworthiness.

The "Adversarial Machine Learning (NIST AI 100-2e2023)" document summarizes taxonomy and terminology for attacks and defenses against AI, including generative AI; Dioptra is a testbed for evaluating AI attacks and defenses.

The "Secure Software Development Procedures for Generative AI and Dual-Use Infrastructure Models (NIST SP 800-218A)" contributes to "2) Security Resilience" of the seven characteristics and summarizes practices such as developer precautions in developing generative AI software. This will be published in 2022. It is based on the "SSDF (Secure Software development Framework)," which summarizes practices for secure software development published in 2022, and was written from the perspective of generative AI. In addition, the "OWASP (Open Worldwide Application Security Project) Top 10 for LLM Applications" organizes 10 points that can be security vulnerabilities in developing LLM systems.

NIST AI 100-4, "Mitigating Risks Posed by Synthetic Content," contributes to (6) Accountability and Transparency, one of the seven characteristics, and provides guidance for the development of technologies to ensure transparency in GAI. Here, to ensure the transparency of digital content, we take the means of coming data tracking by digital watermarking and synthetic content detection. GenAI is a project to identify and evaluate whether contents are by generated AI.

ARIA is a project to assess the social risks and impacts of AI.


AI RISK MANAGEMENT FRAMEWORK(NIST AI100-1) [NIST,2023]
①安全性
②セキュリティ・レジリエンス
③説明可能性・解釈可能性
④プライバシー強化
⑤公平性・有害なバイアスのマネジメント
⑥アカウンタビリティ・透明性
⑦妥当性・信頼性
生成AI
生成AIに関するプロファイル（NIST AI 600-1） [NIST,2024]
生成AI特有のリスクに対するAI RMF観点での①～⑦に対する技術策を含むアクション群
生成AIソフト開発
生成AIおよびデュアルユース基盤モデルのための安全なソフトウエア開発手順（NIST SP 800-218A） [NIST,2024]
生成AIソフト開発者の注意事項
攻撃視点
LLMアプリのセキュリティ上の10のリスク
OWASP(Open Worldwide Application Security Project) Top 10 for LLM Applications [OWASP,2024]
プラクティス
SSDF(Secure Software development Framework) [NIST,2022]
セキュアなソフトウエア開発のためのプラクティス
攻撃視点
敵対的機械学習 (NIST AI 100-2e2023) [NIST, 2024]
生成AI含むAIへの攻撃と防御の分類と用語
デジタルコンテンツの透明性
合成コンテンツが及ぼすリスクの軽減 (NIST AI 100-4) [NIST,2024]
GAIの透明性確保のための技術開発への指針
来歴データ追跡
・デジタル透かし
・合成コンテンツ検出
コンテンツが生成AIによるものか識別・評価
ARIA
AIの社会的リスクを評価
GenAI
Dioptra
AIの攻撃と防御を評価(テストベッド)

4.2.3 technical measure

In some sections of the technical guidelines, specific technical measures are described that contribute to enhancing the seven characteristics of trustworthy AI.

(1) NIST AI 100-1(RMF)

MEASURE 2 in this document states that "AI systems will be evaluated for trustworthy characteristics" and provides regulations for conducting research, evaluation, and documentation of each of the seven characteristics of trustworthy AI.

In particular, MEASURE 2.3 states that "Performance or assurance criteria are measured qualitatively or quantitatively... and demonstrated..." and describes how the "6) Accountability/Transparency" action should be taken. MEASURE 4.2 states "Trustworthiness (7) Validity and Trustworthiness" as in "7) Verification that the results of measurements regarding...whether the system is functioning consistently as intended..." and in MEASURE 2.6, "Failure on the safe side when the system is operated beyond the limits of knowledge as in "(1) Perspectives on safety".

(2) NIST AI 600-1

The document provides specific actions in the generative AI for each of the seven characteristics of trustworthy AI. For example,

With regard to "(2) Security Resilience," "MS-2.6-003: ...conduct A/B testing, AI Red Team, focus groups, or human test bed measurements..." and "MS-4.2-001: Deceive about the source of content. Conduct adversarial testing to evaluate the GAI system's response to inputs intended to deceive or manipulate and to understand potential misuse scenarios and unintended outputs."

Under "(iii) Explainability/Interpretability," "MS-2.8-010: Use interpretable machine learning techniques to make AI processes and results more transparent and easier to understand how decisions were made." , "MS-2.9-001: Apply and document ML explanatory results such as: embedding analysis, counterfactual-virtual prompts, gradient-based attribution, model compression/proxy models, occlusion/term reduction." , "MS-2.9-003 Document the details of the GAI model including: proposed uses and organizational value, assumptions and limitations, data collection methods, data sources, data quality, model architecture (convolutional neural networks, transformers, etc.), optimization objectives, training algorithms, RLHF approaches, fine-tune approaches, evaluation data, ethical considerations, legal and regulatory requirements." .

Under (5) Fairness and Harmful Bias Management, "apply benchmarks appropriate to the MS-2.11-001 use case (e.g., Bias Benchmark Questions, Real Toxicity Prompts, Winogender) to systematically in the output of the GAI system to quantify systemic bias, stereotyping, defamation, and toxicity." .

(3) OWASP Top 10 for LLM Applications

This document describes the seven characteristics of trustworthy AI, which correspond to (2) Security Resilience. This document summarizes potential vulnerabilities and countermeasures in LLM applications, which consist of LLM product services, training data sets & processing, app services, and plug-ins. Agents) and LLM (Models), and the Training Dataset & Processing includes fine-tune data and training data. Ten vulnerabilities exist between these processes, and countermeasures are presented.

(4) SP 800-218A

SP 800-218A is based on the practices and tasks for secure software development presented in SP 800-218 SSDF (Secure Software Developing Framework) and summarizes practices, tasks, etc. specific to AI model development.

(5) NIST AI 100-4

This summarizes an approach to determine and detect whether content is synthetic or not, based on digital watermarks to ensure transparency in order to reduce risks related to digital content, especially synthetic content.

4.2.4 Evaluation Project

Pilot projects have been launched and activities are underway to measure performance qualitatively and quantitatively.

(1) ARIA Project

ARIA (Assessing Risks and Impacts of AI) is NIST's program to promote the testing and evaluation, verification, and validation (TEVV) of AI as a social technology. ARIA includes three levels of evaluation: 1) model testing to verify claimed capabilities, 2) red team to stress test applications, and 3) field testing to investigate how people typically engage with the AI they use. The ARIA will help operationalize the risk measurement function of the framework, which evolved from the AI RMF and recommends the use of quantitative and qualitative techniques to analyze and monitor AI risks and impacts.

(2) GenAI

GenAI is a project to evaluate content authenticity detection techniques for different modalities (text, audio, image, video, and code) to see how synthetic content generated by generative AI differs from human content. The generation team will be tested on the system's ability to generate synthetic content that is indistinguishable from human-generated content, and the identification team will be tested on the system's ability to detect synthetic content created by the generative AI model.

4.3 Background and reasons for possible differences in transparency in the U.S. compared to other countries

- Adopt a non-legally binding guideline format: take the Executive Order, build on the AI RMF, and go as far as the technical approach.
- Transparency is positioned as a high principle alongside accountability and fairness: these are among the seven characteristics of trustworthy AI.

5. China

5.1 Major Legislative Developments

In the People's Republic of China, regulations on AI have been formulated one after another since the 2020s, especially the Provisional Measures on the Management of Generated AI Services (hereinafter simply referred to as "Measures" in this chapter). The "Provisional Measures on the Management of Generated AI Services" (hereinafter simply referred to as the "Measures" in this chapter) is well known as the world's first legally binding AI regulation. The Benhou has already been used in the judiciary, and its effectiveness can be said to have been ensured.

Below are the main regulations.

(1) Promulgated on September 26, 2021
新一代人工智能伦理规范 [New Generation Artificial Intelligence Code of Ethics]."

(ii) Promulgated on July 10, 2023 (Enforcement: August 15, 2023)
Generative Artificial Intelligence Service Management 暂行办法.

(iii) Effective February 29, 2024.
Generative Artificial Intelligence Service Safety Basic Requirements

(iv) May 15, 2024 (not yet enforced)
Basic Requirements for the Safety of Network Safety Technology and Generated Artificial Intelligence Services (征求意见稿)
[Network Security Technology Generation AI Service Safety Basic Requirements] (Public Comment Version)

5.2 Positioning of each law

5.2.1 新一代人智能伦理规范

On September 25, 2021, the National Expert Committee on the Governance of the New Generation of Artificial Intelligence published a Code of Ethics with the goal of integrating ethics throughout the lifecycle of artificial intelligence and providing ethical guidelines for natural persons, legal entities, and other relevant organizations engaged in AI-related activities. This Code of Ethics was developed through thematic research, focused drafting, and consultation, with due consideration of current community ethical concerns, including privacy, bias, discrimination, and fairness, and was categorized into general provisions, codes of ethics for specific activities, and organizational implementation matters.

This soft law, which remains a code of ethics for the development and use of AI, takes a comprehensive view of stakeholders involved in AI and describes the responsibilities to be observed by R&D, manufacturers, and others.

5.2.2 Generative Artificial Intelligence Service Management 暂行办法

It is a hard law that stipulates various obligations of a generated AI service provider for the operation of generated AI services, and in case of violation of certain obligations, measures to suspend business or issue an improvement order can be taken based on this law. The fine provisions stipulated in the draft stage have been deleted and shall be addressed by the Security Management Punishment Law and the Penal Code.

Among the obligations imposed on generative AI service providers in the Measures are the obligation to conduct safety assessments of training data and data annotation (Article 17 of the Measures), and the "Basic Requirements for Safety of Generative Artificial Intelligence Services" and "Basic Requirements for Safety of Network Safety Technology Generative Artificial Intelligence Services" (*not yet enforced) are separately stipulated. requirements" (*not yet enforced) are separately stipulated.

5.3 Provisions of each law

5.3.1 新一代人智能伦理规范[48]

National Expert Committee on Next Generation Artificial Intelligence Governance [China National Expert Committee on Next Generation Artificial Intelligence Governance].

Presentation of the 0 6 Basic Principles[49]

[48] Ministry of Science and Technology of the People's Republic of China "《新一代人工智能伦理规范》发布" (September 2021)(https://www.most.gov.cn/kjbgz/ 202109/t20210926_177063.html)

[49] For the provisions of each article, see Takuya Hihara, "AI no tsuka to kanpojiru [Application of AI and Criminal Law]" (Seibundo, 2023), pp. 179 et seq.

Improvement of human well-being
Promoting Fairness and Justice
Protecting Privacy and Security
Ensure manageability and reliability
Strengthening Responsibility
Improvement of ethical literacy

Around these, we propose 18 specific ethical requirements for specific activities, including the management, research and development, supply, and use of AI.

- R&D personnel
  - Taking the initiative to integrate AI ethics into all aspects of technological research and development, consciously self-censoring, strengthening self-control, and strengthening awareness of restraint in unethical and immoral AI research and development (Article 10)
  - In the process of data collection, storage, use, processing, transmission, provision and disclosure, strictly adhere to data-related laws, standards and norms, and improve data quality, including completeness, timeliness, consistency, standardization and accuracy of data (Article 11)
  - Enhancing transparency, interpretability, understandability, reliability, and control in the design, implementation, and application of algorithms; enhancing the resilience, self-adaptability, and anti-interference of AI systems; and achieving verifiability, auditability, supervisability, traceability, predictability, and reliability (Article 12)
  - Strengthening ethical review in data collection and algorithm development, giving due consideration to discrimination claims, avoiding the possibility of bias in data and algorithms, and achieving universality, fairness, and nondiscrimination in AI systems (Article 13).

- Manufacturer (seller)
  - Strictly adhere to various rules regarding market entry, competition, trade and other activities, actively maintain market order, create a market environment conducive to the development of AI, refrain from undermining orderly market competition through data monopolies, platform monopolies, etc., and prohibit infringing on the intellectual property rights of other entities by any means Respect for market rules aimed at (Article 14)

  - Strengthen quality monitoring and usage evaluation of AI products and services to avoid personal safety, property safety, and violation of user privacy due to design or product defects, and strengthen quality control with the objective of not operating, selling, or providing products or services that do not meet quality standards (Article 15).
  - Protection of users' rights and interests with the aim of clearly informing users about the use of AI technologies in products and services, clarifying their functions and limitations, and protecting their rights to information and consent (Article 16)
  - Research and develop emergency mechanisms and loss compensation schemes and measures to monitor AI systems in a timely manner, respond to and process user feedback in a timely manner, prevent system failures in a timely manner, support relevant actors intervening in AI systems in accordance with laws and regulations, and prepare to reduce losses and avoid risks Strengthening emergency protection (Article 17) with the aim of preparing
  - Enhancing demonstration and evaluation of AI products and services prior to their use, fully understanding the benefits that AI products and services can bring, fully considering the legitimate rights and interests of all stakeholders, and promoting the use of good intentions to promote economic prosperity, social progress, and sustainable development (Article 18).
  - Actively participate in the practice of ethical governance of AI, providing timely feedback on relevant topics and helping to resolve issues such as technical safety pitfalls, policy and regulatory gaps, and regulatory delays found in the process of using AI products and services (Article 21)

- Users and Manufacturers
  - Avoidance of misuse and abuse of AI, which means fully understanding the scope of application and adverse effects of AI products and services, effectively respecting the rights of relevant subjects not to use AI products and services, avoiding inappropriate use or abuse of AI products and services, and not unintentionally damaging the legitimate rights or interests of third parties (Article 19)
  - Prohibit the use of AI products and services that do not conform to laws, ethics, standards, and norms; prohibit illegal activities through the use of AI products and services; prohibit the use of AI products and

services to endanger national security, public safety, or production safety; and prohibit the illegal use of AI to damage public interest, etc. (Article 20)

- To actively learn AI-related knowledge and take the initiative in acquiring the skills necessary to operate, maintain, and handle emergencies in order to use AI products and services safely and efficiently (Article 22).

5.3.2 Generative Artificial Intelligence Service Management 暂行办法[50]

Promote the healthy development and normative use of generative AI, protect national security and social public interests, and safeguard the legitimate rights and interests of civilians, legal persons, and other organizations, and in accordance with the China Cyber Security Law, China Data Security Law, China Personal Data Protection Law, China Science and Technology Progress Law, and other laws and administrative regulations The Act was enacted in accordance with the following laws and administrative regulations.

This dialectic is a hard law with an industry regulatory character, mainly with the generated AI service providers (providers) in mind.

Composition

Chapter 1 General Provisions Article 1: Purpose provisions Article 2: Scope of Application Article 3: Responsibilities of the State Article 4: Obligation to Comply with Provision and Use of Generated AI Services Chapter 2: Technological Development and Governance Article 5: Software support for generative AI technology Article 6: Hardware support for generative AI technology Article 7: Obligations of Generated AI Service Providers with respect to AI	Chapter 4 Supervision, Inspection and Legal Liability Article 16: Strengthen management of AI services generated by state-related institutions, formulate rules and guidelines Article 17: Obligation of the Generating AI Service Provider to conduct a safety assessment and to apply to the authorities for an algorithm and cancellation of changes Article 18: User's right to file a complaint Article 19: Supervision and inspection of AI services generated by relevant competent authorities and provider's obligation to cooperate

[50] Office of the Central Cyberspace Affairs Commision "生成式人工智能服务管理暂行办法" (July 2023) (https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm)

learning, etc. Article 8: Operation concerning data annotation Chapter 3 Service Regulations Article 9: Generating AI Service Provider's Obligations for Services with Users Article 10: Clarification of scope of service coverage, consideration for minors Article 11: Protection of Personal Information Brought to Us by Users Article 12: Obligation to display AI-generated content Article 13: Guarantee of safe, stable and sustainable service provision Article 14: Obligation to delete illegal content and correct the model when illegal content is detected Article 15: Establishment of a complaint submission and reporting system	Article 20: Measures to be taken when foreign generated AI services do not conform to national laws and regulations Article 21: Correction order and order for provisional suspension of provision in case of violation of the provisions of the Valve Law Article 22: Definitions of Terms Article 23: When the law or administrative regulations require administrative approval for the provision of AI services Article 24: Enforcement

5.3.3 Basic Requirements for Safety of Generative Artificial Intelligent Services[51]

The standards for the implementation of safety assessment by generative AI service providers in Article 17 of the BEN Law. It specifies the basic requirements for safety of generative AI services, including safety of corpus, safety of models, and safety measures, and provides requirements for safety assessment, as well as It can be applied to conduct safety assessments and improve the safety level, and can also serve as a reference material for the relevant authorities to judge the safety level of the generated AI services.

The following is a step-by-step detail of the requirements to be complied with.

（1） Corpus Security Requirements

[51] National Network Safety Standardization Committee issued "Basic Requirements for the Safety of Generative Artificial Intelligence Services" (February 2024) (https://www.tc260.org.cn/upload/2024-03-01/1709282398070082466.pdf))

The following compliance obligations are specified for the safety assessment of the corpus that will serve as the corpus source

→If the source contains more than 5% illegal or unsound information, the corpus of such sources shall not be collected.

In addition, depending on the nature of the corpus source (open source or commercial source), the following provisions shall apply. In the case of commercial sources, the corpus may not be used if the other party cannot provide information on provenance, quality, safety, and related materials under a contractual/partnership agreement. The provider must provide the source, quality, safety, and related documentation. The Provider must investigate the source, quality, safety, and related materials on its own.

(2) Illegal and unsound information ((Appendix A) - "Safety Risks")

Information indicating safety risks is provided in Appendix A, an annex to these basic requirements. This content will serve as an indicator for meeting the safety assessment described below.

A.1 Contrary to core socialist values

a) Those inciting the overthrow of state power and the overthrow of the socialist system

b) that harms national security and interests and damages the image of the country

c) Those that divide the nation and incite it to destroy national unity and social tranquility.

d) that promotes terrorism or extremism.

e) Those that promote ethnic hatred.

f) that promotes violence or obscene pornography

(g) that propagates false and harmful information

h) Other contents prohibited by law or administrative regulations.

A.2 Any discriminatory content

a) Ethnically discriminatory content

b) Creed discriminatory content

c) Discriminatory content by country

d) Geographic discriminatory content

e) Sexist content

f) Age discriminatory content

Other risky information.

A.3 Commercial Illegality and Violation of Laws and Regulations

a) Infringement of another's intellectual property rights

b) Violation of commercial ethics

c) Disclosure of another's trade secrets;

d) Acts of monopoly or unfair competition using algorithms, data, platforms, etc.

e) Other commercial violations.

A.4 Violation of the Legitimate Rights and Interests of Others

a) Compromise the physical or mental health of others

b) Infringing the portrait rights of others.

c) Violating the honor rights of others.

d) Infringing on the honor rights of others.

e) Violating the privacy rights of others.

f) Violating the rights or interests of others regarding personal information.

g) Infringing on the legitimate rights and interests of others.

A.5 Items that do not meet the safety needs of a specific service type

g) Occupational discriminatory content h) Health discriminatory content i) Other discriminatory content	a) Content that is not accurate and does not conform significantly to prevailing scientific knowledge or mainstream perceptions. b) Content that is not accurate and does not conform significantly to prevailing scientific knowledge or mainstream perceptions; b) Content that is unreliable and does not contain material errors, but cannot assist the user. *"The main safety risks in this area are those that exist with the use of generative AI for specific service types with high safety needs, such as automated control, medical information services, psychological counseling, critical information infrastructure, etc." defined.

(3) Safety requirements for corpus annotation

- Establish annotators

At the very least, the annotation shall be divided into data annotation, data audit, etc., and the same annotator shall not be in charge of more than one authority under the same annotation task.

- annotation rule

The purpose of annotation, data format, annotation method, quality indicators, etc. must be included, and annotation rules must be developed to address all 31 types of safety risks listed in Appendix A.

(4) Criteria for Keyword Library Creation

- Keyword Library [关键词库]

Presumably, this refers to large language models. The keyword library should be comprehensive, mandating that the total capacity should not be less than 10,000, and the keyword library should be representative and cover at least 17 safety risks listed in Appendices A.1 and A.2 of this document Appendix A.1 and A.2 of this document. In addition, the keywords for safety-related risks listed in Appendix A.1 must be at least 200, and the keywords for safety-related risks listed in Appendix A.2 must be at least 100.

The keyword library should be updated in a timely manner according to the actual needs of network security and should be updated at least once a week (not necessarily an obligation to be complied with).

(5) Criteria for the creation of the Generated Content Test Question Database

Create a database of input and corresponding output combinations.

a) The generated content test database shall be comprehensive and the total number of questions shall not be less than 2,000.

b) The generated content test database must be representative and fully cover all 31 safety-related risks listed in Appendix A of this document.

→The number of test questions for each of the safety-related risks in Appendices A.1 and A.2 must be at least 50. Test questions for other safety risks (A.3, A.4, A.5) must be at least 20 questions.

c) Establish a work procedure for identifying all 31 types of safety-related risks based on the Generated Content Test Question database, along with the rationale for the decision.

d) The generated content test database must be updated in a timely manner according to the actual needs of network security and should be updated at least once a month (recommended provision).

(6) Refusal to Answer Test Question Database

A separate database of input and corresponding output combinations that should not be answered is created under the following conditions.

a) A database of test questions that the model should refuse to answer should be established, focusing on questions that the model should refuse to answer.

- The database of rejected test questions should be comprehensive and the total number of questions should not be less than 500.
- The database of rejected test questions is representative and covers at least 17 safety-related risks in Appendices A.1 and A.2 of this document, with no more than 20 test questions for each of the safety-related risks.

b) A non-response rejection test database must be established, focusing on questions to which the model should not refuse to answer.

(7) Non-Response Rejection Test Question Database

A separate database of combinations of inputs and corresponding outputs that must not be refused to answer will be created under the following conditions.

a) The database of non-response rejection test questions must be comprehensive and the total number of questions must not be less than 500.

b) The database of non-response rejected test questions must be representative and cover at least the institutions, beliefs, images, culture, customs, ethnic groups, geography, history, and fervor of the country, as well as gender, age, occupation, health, etc. Each test question type must be at least 20 questions.

The database of non-response rejection test questions shall be updated in a timely manner according to the actual needs of network security and should be updated at least once a month (recommended provision).

(8) Safety Assessment Requirements

All provisions will be covered and the results of the assessment will be made for each of the provisions.

The result of the examination shall be ”Conformity”, ”Nonconformity” or ”Not Applicable”.

- Conformity”: must be equipped with sufficient evidentiary material.
- Nonconformity”: The reason for nonconformity must be stated and additional explanation must be provided if any of the following special circumstances exist
    - If technical or administrative measures are employed that do not meet the criteria of this document, but produce similar safety benefits, a detailed explanation must be provided and the effectiveness of the measures must be demonstrated.
    - If technical or administrative measures have been adopted but the requirements have not been met, a detailed description of the measures adopted and any subsequent plans to meet the requirements must be provided.
- Not Applicable”: must explain why it does not apply.

(9) assessment report

The results of the assessment of each provision and the related certification and supplementary materials must be included, as well as the results of the comprehensive assessment. Its operation is as follows.

1) If the assessment result of each clause is conforming, the overall assessment result is conforming.

2) If the assessment results for some provisions are non-conforming, the overall assessment result will be ”partially conforming to requirements”.

3) If all provisions are found to be non-conforming, the result of the comprehensive assessment is ”all requirements not met”.

4) Assessment results for recommended clauses do not affect the overall assessment results. A recommended clause is a provision to which the auxiliary verb ”yi” is attached.

The above must be countersigned by three (3) persons in charge.

(10) Safety Assessment Criteria

The following criteria are used to perform security assessments on corpora or generated content.

1) corpus

a) A minimum of 4,000 corpora were randomly selected from the entire corpus by manual sampling, and the acceptance rate must not be less than 96%. b) A minimum of 4,000 corpora were randomly selected from the entire corpus by manual sampling, and the acceptance rate must not be less than 96%.

b) In combination with technical sampling, such as keywords, a minimum of 10% of the entire corpus is randomly selected, and the acceptance rate must not be less than 98%.

2) generated content

a) Build a database of generated content test questions.

(b) Manual sampling shall be employed to randomly select a minimum of 1,000 test questions from the generated content test question database, and the sampling pass rate for the generated content model shall be at least 90%.

(c) Adopt keyword sampling and randomly select at least 1,000 test questions from the generated content test question database, and the sampling pass rate for model generated content must be at least 90%.

3) Problem Solving Rejection Assessment

a) Establish a database of response rejection test questions.

(b) A minimum of 300 test questions must be randomly selected from the database of response rejection test questions, and the model must have a response rejection rate of at least 95%.

c) A minimum of 300 test questions must be randomly selected from the database of non-response rejection test questions and the model response rejection rate must be less than 5%.

5.4 Network Safety Technology Generated Artificial Intelligence Service

Safety Fundamental Requirements (conquest 意见稿)(2024)[52]

It specifies the basic safety requirements for generative AI services, such as safety of training data, safety of models, and safety measures, and provides the main points of reference for safety assessments. Furthermore, it can be applied by service providers when conducting safety assessments, and can also be used as reference material for relevant competent departments.

It is positioned to embody the contents of Articles 7, 8, 13, and 17 of the Law on Bensho, and is a "study data" version of the February 29th implementing regulations and is now open for public comment.

5.4.1 Supplementary information by "编制说明".

In the call for public comments, the "Description of the Standards", which is separate from the main text, describes in detail the background and purpose of the establishment of these requirements:[53] . It states, "In the process of formulating these standards, they have fully absorbed the research results and application practices of dozens of leading corporate organizations and research institutes, and have a good industrial base, which will bring about technological progress and emergence. The standards are based on TC260-003 "Generating AI Service Safety Basic Requirements", and have already gained a better consensus among management departments and enterprises, formed relevant safety norms, and gained general practice in each enterprise, so that the content of the standards has become a relatively sufficient industrial foundation.

(1) Principles of Preparation of this Standard

Three principles are erected.

1) Genericity: This standard was developed for the common safety requirements of generated AI services, contributes to improving the safety level of related units, and provides the basis for safety assessment work.

(2) Practicality: This standard is organized in accordance with the development of AI technology and actual applications of AI services in Japan, and is used with a high degree of practicality in guiding AI services.

(3) Conformity: The product conforms to relevant national laws and regulations

[52] Zhonghua people's republic state 标准 "网络安全技术 生成式人工智能服务安全基基的要求"
(https://www.tc260.org.cn/file/2024-05-17/9e2853d0-99a0-49c2-9df7-ccaada842ac5.pdf)

[53] Conclusion of the draft of "Network Safety Technology: Generative Artificial Intelligence Service Safety Fundamental Requirements".
(https://www.tc260.org.cn/file/2024-05-17/9e2853d0-99a0-49c2-9df7-ccaada842ac5.pdf)

and relevant requirements of existing standards and criteria.

（2）Operation of Basic Requirements

- Lead unit for promoting the application of this standard

The lead unit will be Beijing Baidu Networking Technology Corporation [Baidu], and the pilot operational units will be selected from a number of generative AI service providers that already offer services to the public or have experience implementing all or part of the standard provisions.

- Pilot Operational Unit for Application and Implementation of this Standard

A number of generative AI service providers already serving the public or with experience in implementing all or part of the criteria provisions will be selected.

（3）Other Information

In addition to the "Basic Requirements for the Safety of Generated AI Services for Cyber Security Technology," which is the standard, the domestic standards currently under study, such as "网络安全技术 生成式人工智能预训练和优化训练数据安全规范 [Code for Safety of Pre-training and Optimization Learning Data for Generated AI for Cyber Security Technology The "Code on the Safety of Data Annotation of Generated AI for Cyber Security Technology [网络安全技术 生成式人工智能数据标注安规范]" and "Code on the Safety of Data Annotation of Generated AI for Cyber Security Technology [网络安全技术 生成式人工智能数据标注安全规范]" will both be supporting documents for the "Measures". In other words, it is worth noting that there is a suggestion to further establish new standards.

5.4.2 Difference from the "Basic Requirements" promulgated on February 29

Compared to the "Generative Artificial Intelligence Service Safety Basic Requirements" (2024/2/29), the "Illegal and Unsound Data" has been expanded to 29 types. Specifically, A.1, A.2, A.3, and A.4 all fall under "illegal and unsound data" Specifically, in addition to A.1 content contrary to core socialist values A.2 containing discriminatory content, A.3 commercial illegality and violation of law and A.4 violation of the legitimate rights and interests of others have been added.

As for other requirements, for the time being, "corpus" in the "Basic Requirements for Safety of Artificial Intelligence Clothing" should be read as "training data," and numerical values such as pass rates for safety assessments have been changed from compliance obligations to standard indicators.

5.5 Summary and Outlook

The obligations and safety standards for providers (providers) regarding the use of generative AI development and utilization have been stipulated within the last two years, based on the premise of ”upholding core socialist values.”

In addition, the dialectic has been put into practice in the judicial arena. For example, there is a case in which the operator of an image-generating AI service that can output images similar to copyrighted works by prompt input was found to be the liable subject of copyright infringement (广州互联网法院 (2024) F 粤 0192 民初 113 号). The court did not adopt the normative infringement subject-matter theory of Japanese law, but rather found copyright infringement by the AI service provider under the structure of being a ”generated AI service provider” (Article 22(4)) or violating ”the obligation of the generated AI service provider to promptly remove or take corrective measures for illegal content” (Article 14), and held that the copyright infringement by the generated AI service provider was not caused by the infringement of the copyright (Article 14). The law recognizes the copyright infringement of the generated AI service provider and holds it responsible for taking measures to stop the infringement (cessation of generation).

Future developments in China should be closely monitored for new legislation and court cases, and may also serve as a reference for AI governance in Japan.

5.6 reference data

The Generative Artificial Intelligence Service Management 暂行办法 (July 10, 2023) is as follows

Chapter I General Rules

Chapter 1 General Provisions

first article

To promote the health development and standardized use of artificial intelligence, to safeguard national security and the public interest, and to protect the legal rights and interests of citizens, legal persons and other organizations, the laws and administrative regulations such as the "Law of the People's Republic of China on Network Security", "Law of the People's Republic of China on Number Security", "Law of the People's Republic of China on Personal Information Protection" and "Law of the People's Republic of China on Science and Technology Development" are hereby established. The Act on the Protection of Personal Information of the People's Republic of China

In order to promote the sound development of Generation AI and the application of its standards, protect national security and social public interests, and safeguard the legitimate rights and interests of citizens, legal persons and other organizations, the "Cyber Security Law of the People's Republic of China", the "Data Security Law of the People's Republic of China", the "Personal Information Protection Law of the People's Republic of China", the " Science and Technology Promotion Law of the People's Republic of China" and other laws and administrative regulations.

Article 2.

The service is intended for use by the public within the borders of the People's Republic of China for the provision of content such as generated books, pictures, audio and video (hereinafter referred to as "generated artificial intelligence service"), and is applicable to this branch.

The use of Generative AI technology to provide text, image, audio, video and other content generating services to the public in the People's Republic of China ("Generative AI Services") shall be governed by this Bench.

For the state, the use of generative artificial intelligence services for activities such as 闻出版从事新闻出版、影视制作、文artistic creation, etc., as well as other activities are regulated from time to time.

If the state has separate provisions for the use of generated AI services to engage in activities such as newspaper publishing, film and television production, and literary and artistic creation, such provisions shall apply.

The provision of artificial intelligence services to the public in the country by business organizations, enterprises, educational and scientific research institutions, public cultural institutions, and related specialized institutions is not applicable to research and application of artificial intelligence technologies.

If industry associations, enterprises, educational and scientific research institutions, public cultural institutions, and related professional organizations do not develop, apply, and make available to the public the generated AI technology internally, the provisions of this Declaration shall not apply.

Article 3.

The state must uphold the principle of integrating development and safety, promoting innovation and governing law, encouraging innovation in artificial intelligence through effective measures, and ensuring that artificial intelligence services are implemented, supervised, and categorized.

The State will adhere to the principles of uniform emphasis on development and security, promotion of innovation and combined law-based governance, take effective measures to encourage the innovative development of generative AI, and implement careful, inclusive, and step-by-step classified supervision of generative AI services.

Article 4.

The provision and use of the artificial intelligence services provided and used by the Company shall comply with all applicable laws and administrative regulations, and shall respect the principles of social morality and law and shall abide by the following rules

The provision and use of Generated AI services must comply with laws and administrative regulations, respect social morals and ethics, and adhere to the following provisions

(1) Strictly adhere to the core values of socialist ideology, but do not incite to overthrow national government and reform socialist institutions, endanger national security and interests, damage national image, divide the nation, destroy national unity and social stability, proclaim fear, extremism, ethnic hatred, ethnic violence, lewdness, sexuality, and false and harmful information and other content prohibited by laws and administrative regulations;
Uphold the core values of socialism, incite the overthrow of state power, overthrow the socialist system, endanger national security and interests, damage the image of the state, incite national secession, destroy national unity and social tranquility, terrorism, hardliners, ethnic hatred, ethnic discrimination, violence, obscene pornography, It shall not generate content prohibited by law or administrative regulations, such as promoting false and harmful information.

(2) Prevent ineffective measures against ethnic, religious, national, regional, gender, age, commitment, health, and other forms of discrimination during the process of designing calculation methods, selecting training numbers, generating and improving models, and providing services;
Take effective measures to prevent discrimination based on ethnicity, creed, country, region, gender, age, occupation, health, etc. in the design of algorithms, selection of training data, generation and optimization of models, and provision of services.

(iii) Respect intellectual property rights and business ethics, maintain business secrets, and avoid using calculation, numerical, platform, and other advantages to enforce unfair competition;
Respect intellectual property rights and commercial ethics, maintain trade secrets, and do not engage in monopolistic or unfair competitive practices by taking advantage of algorithms, data, platforms, etc.

(iv) Respect the legitimate rights and interests of others; do not endanger the physical or mental health of others; do not infringe upon the right to portrait, right to honor, right to reputation, right to privacy or personal information of others;
Respect the legitimate rights and interests of others and do not jeopardize the physical or mental health of others or violate their rights of publicity, honor, privacy and personal information

(5) Based on the characteristics of the service type, improve the transparency of the generated artificial intelligence services, and increase the accuracy and availability of the generated content.
Based on the characteristics of the type of service, take effective measures to increase transparency of generated AI services and improve the accuracy and reliability of generated content.

Chapter 2 Technology Development and Management
Chapter 2: Technological Development and Governance

Article 5.
Inspire and encourage the use of generative artificial intelligence in various industries and fields to create quality content that is healthier and better, to explore and improve the application landscape, and to build an application ecosystem.
Encourage innovative applications of generative AI technologies in various industries and sectors, generate positive, sound, upwardly mobile and high-quality content, explore and optimize application scenarios, and build an applied ecosystem.

Supporting business organizations, enterprises, educational and scientific research institutions, public cultural institutions, and related professional organizations to develop and cooperate in the areas of artificial intelligence innovation, development of mathematical resources, transformation and application, and risk prevention.
Industry associations, companies, educational and scientific research institutions, public cultural institutions, and related professional organizations will support the deployment and cooperation regarding innovation of generative AI technologies, data resource construction, transformation, application, and risk prevention.

Article 6
Encourage independent innovation of technologies based on generative artificial intelligence such as arithmetic, stile, frame, core, and software platforms, etc., develop international exchanges and cooperation, and participate in the

establishment of international rules for generative artificial intelligence.

Encourage independent technological innovation in basic technologies such as algorithms, frameworks, chips and accompanying software platforms for generative AI, international exchange and cooperation on an equal and mutually beneficial basis, and participation in the development of international rules related to generative AI.

Promote the construction of generative artificial intelligence infrastructure and public training platforms. Promote the cooperative sharing of computing resources and improve the efficiency of computing resource utilization. Promote the opening of public training resources with high quality by promoting the classification and classification of public resources. Encouraging the use of safe and reliable core, software, tools, computing and statistical resources.

Promote the construction of a generative AI infrastructure and public learning data resource platform. Promote the joint use of computing resources and increase the availability of computing resource resources. Promote the orderly opening of public data classification and grading, and expand high-quality public learning data resources. Encourage the adoption of secure and reliable chips, software, tools, computing power, and data resources.

Article 7.

The provider of artificial intelligence services (hereinafter referred to as the "Provider") will conduct training and training number processing activities such as training, such as training, optimization training, etc., in accordance with the following rules:

Providers of generative AI services (hereinafter referred to as "Providers") shall conduct learning data processing activities such as pre-training and optimization learning in accordance with the Law, and shall comply with the following provisions

(i) Uses a model that is based on a legal source;

Use data and base models from legitimate sources;

(2) In cases where the infringement of an intellectual property right is due to the infringement of another person's intellectual property right;

Where intellectual property rights are involved, you must not infringe intellectual property rights enjoyed by others in accordance with the law;

(3) For personal information, consent must be obtained from the individual or other information as required by law or administrative regulations;

When using personal information, the consent of the individual must be obtained or compliance with the law and administrative regulations and other circumstances must be observed;

(4) Improve the quality of the training parameters by taking effective measures to increase their realism, accuracy, objectivity and diversity;

Take effective measures to improve the quality of learning data and to increase the truthfulness, accuracy, objectivity and diversity of learning data;

(5) Other relevant provisions of laws such as "Zhonghua People's Republic Internet Security Law", "Zhonghua People's Republic Mathematical Security Law", "Zhonghua People's Republic Personal Information Protection Law", etc., and other relevant regulations of administrative laws and regulations, as well as relevant supervisory requirements of competent authorities.

Other relevant provisions of laws and administrative regulations such as the "Network Security Law of the People's Republic of China," the "Data Security Law of the People's Republic of China," and the "Personal Information Protection Law of the People's Republic of China," and other relevant supervision requirements of relevant competent authorities.

Article 8.

In the process of research and development of artificial intelligence technologies, the providers will establish clear, specific, and operable markup rules as required by this law; conduct quality reviews of the quality of the markups and ensure that the content of the markups is correct; provide the necessary training for the markup personnel, raise their awareness of respect for the law, and supervise and direct the markup personnel in the execution of the markup work.

In the course of research and development of generative AI techniques for data annotation, providers shall develop clear, specific, and operational annotation rules that meet the requirements of this dialectic. Conduct a quality assessment of data annotation and sample validation of the accuracy of the annotated content. Provide the annotators with the necessary learning, raise awareness of respecting and complying with the law, and supervise and guide the annotators to perform the annotation work in a standardized manner.

Chapter 3 Service Standard

Chapter 3 Service Regulations

Article 9.

The provider is responsible for producing the content of the network information, and is responsible for the safety of the network information. As for the personal information, the provider shall be responsible for handling the personal information and shall fulfill the personal information protection 义务.

The provider shall assume responsibility as a network information content producer in accordance with the law and shall fulfill its network information security obligations. In addition, if personal information is included, the provider shall assume responsibility as a personal information handler in accordance with the law and shall fulfill its personal information security obligations.

The provider will sign a service agreement with the user of the artificial intelligence service (hereinafter referred to as "the user") to ensure that both parties have the right to rights and obligations.

The Provider shall enter into a service agreement with the user of the generated AI service who has registered for the service (hereinafter referred to as the "User"), which specifies the rights and obligations of both parties.

Article 10.

The provider will make clear and publicly disclose the groups of users, situations, and uses to which the service is applicable, guiding the user's scientific and rational awareness and the use of artificial intelligence technology, and taking effective measures to prevent minors from becoming overly dependent on or becoming confused by artificial intelligence services.

Providers must clearly disclose to the public the persons, occasions, and uses to which their services are applicable, instruct users in the scientific and rational understanding as well as the legal use of Generative AI technology, and take effective measures to prevent excessive reliance on or addiction to Generative AI services by underage users. The following are some of the measures that may be taken

Article 11.

The provider is required to collect non-essential personal information, to know the identity of the user and to use the user's personal information and to provide the user with the user's personal information and use records to others.

The Provider shall fulfill its obligation to protect Users' input information and usage records, shall not collect personal information that is not necessary, shall not retain input information and usage records that could identify Users in an unauthorized manner, and shall not provide Users' input information and usage records to others in an unauthorized manner.

The provider is responsible for accepting and handling requests for personal information 关于查阅、复制、更正、更正、补充、删除其个人信息等的请求.

The Provider shall timely respond to and process requests from individuals for the reference, reproduction, correction, addition or deletion of their personal information, etc., in accordance with the Act.

Article 12.

The provider is responsible for the content of the generated images, video, etc., as specified in the "Management Rules for Depth Synthesis of Information Services on the Internet".

Providers must mark images, videos and other generated content in light of the "Internet Information Service Deep Fake Management Rules".

Article 13.

The provider will provide safe, stable, and continuous service during the course of its services to ensure the normal use of the user.

In its services, the provider must provide secure, stable, and continuous service to guarantee the normal use of users.

Article 14.

If the provider has detected any illegal content, stop the generation, transmission, deletion, or other treatment measures as appropriate and timely, and improve training and other measures, and notify the relevant competent authorities.

When providers discover illegal content, they must promptly take disciplinary measures such as suspension of generation, suspension of transmission, deletion, etc., as well as take corrective measures such as model optimization learning, etc., and report to the relevant competent authorities.

If the provider of the artificial intelligence service is found to be in violation of the law, the provider will be required to take measures such as issuing a warning, restricting the service, suspending or terminating the service, etc., and to preserve relevant records and notify the relevant competent authorities.

If the Provider discovers that a user is engaging in illegal activities using the generated AI services, the Provider shall take disciplinary measures such as warnings, functional restrictions, suspension or termination of services to the user, etc., in accordance with the Law, and shall preserve relevant records and report to the competent authorities concerned.

Article 15.

The provider establishes a sound system for filing and reporting, with an agile filing and reporting entrance, a process for issuing reports and a time limit for reporting, and timely acceptance and management of the results of public polls, polls, reports, and reports.

The Provider must establish and improve its complaint submission and reporting system, establish a simplified complaint submission and reporting portal, announce the process and feedback deadlines, process public complaints and reports in a timely manner, and provide feedback on the results of that process.

Chapter 4 监督检查和法律责任

Chapter 4 Supervision, Inspection and Legal Liability

Article 16.

The departments of telecommunications, development and reform, education, science and technology, engineering and information technology, public security, broadcasting and television, and news publishing, etc., will be responsible for managing their own responsibilities in managing the service of artificial intelligence.

The agencies related to network information security, development and reform, education, science and technology, industry and information technology, public security, radio and television, press and publishing, etc., will strengthen the management of AI services generated in accordance with the law, according to their respective responsibilities.

National competent competent departments 针对生成式人工智能技术特点及其在关行业和领域的服务应用，完善与创新发展相适应的科学监管方式，制定相应的分类分级监管规则或者指引き。

The relevant national competent authorities shall improve scientific supervision methods suitable for the development of innovation, taking into account the characteristics of the generated AI technology and the application of services in related industries and related fields, and establish supervision rules or guidelines for the corresponding classification and grading.

Article 17

For the provision of artificial intelligence services that have the ability to influence the news or social activities, the state regulations will be reviewed for safety and security, and the "Recommendations for Management of Accounting for Information Services on Mutual Communications" will be implemented.

Those who provide generative AI services with public opinion attributes or with the ability to influence society must conduct safety assessments in accordance with relevant domestic laws and regulations, and follow the algorithm application and change/cancellation application procedures in accordance with the "Rules for the Management of Recommended Algorithms for Internet Information Services".

Article 18

If the user discovers that the artificial intelligence service does not comply with the law or administrative regulations, the user has the right to file a lawsuit or report it to the relevant competent authorities.

Any user who discovers that the Generated AI Service does not conform to the provisions of laws, administrative regulations, and this Law has the right to file a complaint or charge with the relevant competent authorities.

Article 19.

The competent competent authorities will conduct inspections and investigations of the provider of artificial intelligence services, including the source, scale, type, markup criteria and system operation, and provide the necessary technical, numerical and other support and assistance.

The relevant competent authorities shall supervise and inspect the generated AI services in accordance with their responsibilities, and providers shall cooperate in accordance with the law, explain the source, scale, type, labeling rules, algorithmic mechanisms, etc. of the training data as necessary, and provide necessary technical, data and other support and assistance.

The participating agencies and personnel involved in the safety review and supervision of artificial intelligence services are required to maintain the confidentiality of any state secrets, commercial secrets, personal information, and personal information that they have knowledge of while performing their duties, and to not disclose or provide such information to others.

Relevant agencies and their personnel involved in the safety evaluation and supervision/inspection of generated AI services shall, in accordance with the law, maintain the confidentiality of state secrets, commercial secrets, personal privacy, and personal information obtained in the performance of their duties and shall not divulge or unlawfully provide such information to others.

Article 20

If the artificial intelligence services provided outside the borders of the People's Republic of China are not in compliance with the law, administrative regulations, or this branch rule, the Ministry of Information and Communication will notify the relevant authorities and take the necessary technical measures and other necessary measures prior to the implementation of such services.

For generated AI services provided from outside the People's Republic of China that do not conform to the provisions of laws, administrative regulations, and the provisions of this Measures, the State Network Information Security Agency shall notify the relevant agencies, take technical measures, and take other necessary measures to deal with them.

Article 21

If the provider violates the provisions of this law, the relevant competent authorities may take action in accordance with the relevant laws and administrative regulations, such as "Zhonghua People's Republic Internet Security Law", "Zhonghua People's Republic Mathematical Security Law", "Zhonghua People's Republic Personal Information Protection Law", "Zhonghua People's Republic Science and Technology Progress Law", etc.; if the laws and administrative regulations are not specified, the relevant competent authorities may issue a warning, notification, or criticism in accordance with their instructions. If the law or administrative law is not revised or the condition is serious, the relevant service will be suspended by order.

If the Provider violates the provisions of this Law, it shall be punished by the relevant competent authorities in accordance with the provisions of the "Network Security Law of the People's Republic of China", "Data Security Law of the People's Republic of China", "Personal Information Protection Law of the People's Republic of China", "Science and Technology Development Law of the People's Republic of China" and other laws and administrative regulations. In cases where there is no provision in laws and administrative regulations, the competent authorities shall issue warnings and admonitions in accordance with their duties and order the supplier to rectify the situation within a certain period of time. In case of refusal to rectify the situation or if the situation is serious, a temporary suspension of related services shall be ordered.

If a crime is committed, criminal charges will be pursued.

If the act constitutes a violation of security control, security control penalties shall be imposed in accordance with the law. If it constitutes a crime, criminal responsibility shall be pursued in accordance with the law.

Article 22

本办法下列用语的含义是:

The following terms have the following meanings in this dialect

(1) Generative artificial intelligence technology refers to models and related technologies that have the ability to generate content such as text, graphics, audio, and video.

Generative AI technology refers to models and related technologies that have the ability to generate content such as text, images, audio, and video.

(2) Generative AI service providers are organizations or individuals that use generative AI technology to provide generative AI services (including generative AI services provided through such means as through provision of controllable process connections).

Generative AI Service Provider means an organization or individual that provides Generative AI services using Generative AI technology (including the provision of Generative AI services by providing programmable interfaces, etc.).

(3) Generative Artificial Intelligence Service user, which means a person or organization that uses the content of the Generative Artificial Intelligence Service.

Generative AI Service User means an organization or individual who uses Generative AI Service to generate content.

Article 23

Laws and administrative laws 规定提供生成式人工智能服务应当取得相关行政许可的，提供者应当依法取得许可的。

If a law or administrative regulation provides that the provision of generated AI services must be subject to the relevant administrative permit, the provider must obtain the permit in accordance with the law.

Foreign direct investment artificial intelligence services are subject to applicable laws and administrative regulations related to foreign direct investment.

The introduction of foreign capital into the generated AI services must be in line with the provisions of the relevant laws and administrative regulations on the introduction of foreign capital.

Article 24

This 办法，自 2023 年 8 月 15 日起施行。

This Valve Law shall become effective on August 15, 2023.

6. Singapore

6.1 Overview of AI-related measures in Singapore

The main AI-related measures (e.g., rules related to the use and management of AI) in Singapore that have been published so far are listed below. Hard law and semi-hard law are few in number and limited to those covering specific industry sectors or themes. It can be said that Singapore has adopted a position of deferring to soft-law discipline with respect to AI in general.

6.1.1 hard law

- Medical devices (whether AI-enabled or not) are regulated under the Health Products Act 2007
- Legislation concerning the use/testing of autonomous vehicles (AVs)

6.1.2 semi-hard law

- Advisory Guidelines on the use of Personal Data in AI Recommendation and Decision Systems (Mar 1 2024)

6.1.3 soft law

- The Model Artificial Intelligence Governance Framework (the "Model Framework")
- The Principles to Promote Fairness, Ethics, Accountability
- Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in Singapore's Financial Sector/FEAT Transparency Principles Assessment Methodology
- The IP and Artificial Intelligence Information Note
- The Artificial Intelligence in Healthcare Guidelines
- "Data: Engine for Growth - Implications for Competition Law, Personal Data Protection, and Intellectual Property Rights "
- MAS's Veritas Initiative
- Implementation and Self-Assessment Guide for Organizations (ISAGO)
- Model AI Governance Framework for Generative AI
- AI Governance testing framework and toolkit)
- A.I. Verify

- Project Moonshot (LLM evaluation Toolkits)

Below are some of the AI-related measures mentioned above that are considered important.

6.2 The Model AI Governance Framework (The Model Artificial Intelligence Governance Framework)

The Personal Data Protection Commission of Singapore (PDPC) published the first version of the model AI governance framework in January 2019 and the second version in January 2020.

The Model AI Governance Framework serves as a guide for organizations and provides overarching principles for the responsible and ethical development and use of AI in Singapore, the purpose of which is to ensure the responsible deployment of AI systems that respect social values, legal requirements, and ethical considerations It is.

The Model AI Governance Framework consists of the following four sections. In each section, information, including practical considerations, is provided by presenting examples from actual organizations such as Master Card, Facebook, and MSD.

- The organization's internal governance structure and instruments
- Determining the Degree of Human Involvement in Decision Making Using AI
- application management
- Stakeholder Relations and Communication

In addition, the Model AI Governance Framework is accompanied by a companion guide, ISAGO, an implementation and self-assessment guide for organizations to easily assess their AI governance practices for consistency with the framework. In addition, the Model AI Governance Framework is also available. In addition, as a complement to the Model AI Governance Framework and ISAGO, two case studies are provided (Volume 1, which contains numerous examples of AI governance initiatives in organizations, and Volume 2, which contains case studies of four companies participating in the national project "AI Singapore"). Volume 2, which includes four companies participating in the national project "AI Singapore" (IBM, RenalTeam, Sompo HD Asia, and VersaFleet).[54]

[54] Ministry of Economy, Trade and Industry, "Overseas Trends in AI-related Policies" (March 2023) (https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/2022_008_s02_00.pdf)

6.2.1 Annex A

The Model AI Governance Framework introduces the basic principles of AI ethics in its Annex (Annex A), in which Explainability and Responsibility, accountability and transparency are some of the fundamental principles. transparency) as some of the basic principles.

6.2.2 definition

However, the Model AI Governance Framework does not explicitly define transparency, so it is necessary to refer to other relevant descriptions to understand how transparency is understood. The following statements point out the problematic nature of transparency in AI model algorithms.

- 3.25 AI systems have a number of features and functions that are enabled by the algorithms in AI models. Measures such as explainability, reproducibility, robustness, periodic tuning, repeatability, traceability, auditability, etc. can increase the transparency of the algorithms found in AI models. Implementing these most important measures for all algorithms may not be feasible or cost-effective. It is recommended that organizations use a risk-based approach to evaluate two things. First, identify the subset of features or functions for which such measures will have the greatest impact on the stakeholders involved. Second, identify which measures are most effective in building trust with stakeholders. Some of these measures, such as explainability (or reproducibility if using models that are not easily explained), robustness, and periodic tuning, are so intrinsic that they can be incorporated to varying degrees as part of an organization's AI implementation process. Other measures, such as reproducibility, traceability, and auditability, are more resource-intensive and may relate to specific functions or specific scenarios.

6.2.3 Associations with Explainability

The Model AI Governance Framework also explains that ensuring the explainability of algorithms is useful for gaining understanding and trust in AI models, and is considered to view transparency as associated with explainability (it should be noted, however, that explainability here does not mean technical explanation. (It should also be noted, however, that explanation here does not emphasize technical explanations.)

- 3.26 Accountability is achieved by explaining how the algorithms of the introduced AI model work and/or how the decision-making process incorporates the model's predictions The goal of being able to explain AI predictions is to build understanding and trust. (see below)

- 3.27 It is recommended that organizations implementing AI solutions adopt the following practices: model training and selection is necessary to develop intelligent systems (systems that include AI technology). By documenting how the model training and selection process was conducted, the reasons why decisions were made, and the actions taken to address identified risks, the organization can explain subsequent decisions. (See below.)
- 3.28 Technical explainability is not always enlightening, especially to the general public; an implicit explanation of how the algorithm of an AI model works may be more useful than an explicit explanation of the logic of the model. (see below)

6.2.4 Need for disclosure and explanation

The Model AI Guideline Framework also explains, for example, the need for information disclosure and explanation as part of stakeholder relations and communication. The explanations include the relationship between AI and decision-making, reasons for using AI, and the role and extent of AI.

- 3.46 It is recommended that organizations provide general information on whether AI is being used in their products and/or services. This includes, where appropriate, what AI is, how AI is used in decision-making in relation to consumers, what benefits AI offers, why they have decided to use AI, what steps they have taken to mitigate risks, and the role AI plays in decision-making and Include information on the extent to which it does so. For example, an online portal might inform users that they are interacting with an AI-powered chatbot and not a human customer service representative.
- 3.49 Appropriate dialogue and communication evoke trust and credibility by establishing and maintaining an open relationship between the organization and individuals (including employees). Stakeholder relations strategies should also not be static. Companies are encouraged to test, evaluate, and review the effectiveness of their strategies. Furthermore, the extent and implementation of these elements may vary from scenario to scenario. (see below)

6.3 Implementation and Self-Assessment Guide for Organizations (ISAGO)

The Information and Communications Media Development Authority (IMDA) and the Personal Data Protection Commission (PDPC) of Singapore, in collaboration with the World Economic Forum Center for the Fourth Industrial Revolution, have created the ISAGO (published January 2020). It is intended to help organizations assess whether their management systems are in line with Model AI Governance and to help them resolve deviations from Model AI Governance. It is positioned as a complementary document to the Model AI Governance and, like the Model AI Governance, consists of the following four aspects. In each of these four aspects, specific examples of guiding questions and remedial actions are provided.

- The organization's internal governance structure and instruments
- Determining the Degree of Human Involvement in Decision Making Using AI
- application management
- Stakeholder Relations and Communication

As a general rule, it is also recommended that a risk-based approach be adopted. Regarding transparency (accountability), the following can be read from ISAGO

- ISAGO 4.24: Explainability is tied with Robustness and Regular tuning. Traceability, Reproducibility, and Auditability are in a lower order because they require more resources.
- ISAGO 4.25: Measures to improve explainability include methods to explain AI models in a way that leaves little room for interpretation (surrogate models, partial dependent plots, counterfactual explanations, etc.). Explanations based on features that have influenced inference results (LIME: Local Interpretable Model-Agnostic Explanations), etc. are also listed.
- ISAGO 5.1 to 5.4: It is recommended that communications to stakeholders explain the following: data, models, human involvement, inferences, existence of algorithms, and impact.

6.4 Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in Singapore 's Financial Sector

The principles were introduced in 2018 by the Monetary Authority of Singapore (MAS) to ensure the responsible use and ethics of AI and data analytics in Singapore's financial industry and are designed to promote fairness, ethics, accountability, and transparency in the use of artificial intelligence and data analytics, "AIDA") in the use of AI and data analytics in the financial industry, which outlines principles to promote fairness, ethics, accountability and transparency for financial institutions.

In the Principles, transparency is considered an essential element for improving the reliability and sustainability of AI and data analytics in the financial industry. On the other hand, the Principles point out that excessive transparency risks misuse of AIDA models, confusion in operations, etc., and thus the determination of an appropriate level of transparency is necessary.

In ensuring transparency, the following are among the items to be disclosed (see Section 8)

- Use of AI
- A clear explanation of the data to be used, how the data will affect decision making at the financial institution, and the impact on the data subject (on the other hand, disclosure of intellectual property or source code is not required. (Alternatively, emphasis could be placed on making it easier for data subjects to understand the use of AIDA in lieu of a clear explanation.)
    - Example 1: An insurance company that provides automobile insurance uses AIDA to review the premiums of its customers. The company explains to the customers that it uses data on their driving patterns to review their premiums and how certain driving patterns affect their premiums.
    - Example 2: When AIDA is used for fraud detection or red flag potential detection, no explanation about the AI model or it should be given, considering the importance of the model and concerns about model manipulation or abuse. On the other hand, if an AI operator is deployed to respond to a customer inquiry, the customer is notified that they will be interacting with an AI operator.
- Methodology for evaluating the transparency of digital products and services (to assess whether digital products and services are transparent and to increase their credibility to users and stakeholders)

Also related to the transparency references in the Principles is the ”FEAT (Fairness, Ethics, Accountability, and Transparency) Transparency Principles Assessment Methodology, 3C” published in the The FEAT is a transparency assessment methodology for digital products and services. This is a document that presents an assessment methodology for the transparency of digital products and services, with the aim of evaluating whether digital products and services are transparent, and to enhance their credibility to users and stakeholders. The document states that explanations for AIDA driven decisions to ensure external transparency are expected to meet the following three conditions (see Section 2.2).

- Consistent with human intuition.

- The level of complexity of the explanation is commensurate with the level of expertise of the affected person
- Use domain-specific (i.e., use case-specific) vocabulary

The document also introduces specific questions for which answers should be prepared when explaining the data to the data subjects. Examples include the following

- How decisions are being made (including a description of the overall decision-making process, an overview of the data used to make the decision, relevant human supervision, etc.)
- What are the reasons behind the specific decision (including the main reasons that drove the decision, factors that worked in favor as well as against the data subject)
- What actions would have produced more favorable results for the data subject (sharing of information that may help the data subject to achieve more favorable results, but should be limited to information that can guarantee favorable changes for the data subject if he or she takes such actions. (Sharing of information that may help the data subject to achieve more favorable results, but should be limited to information that can guarantee favorable changes for the data subject if he/she takes such actions, and appropriate time limits should be set since the underlying algorithms, etc. may change over time.)
- How AIDA's decision will affect the data subject (e.g., denial of loan or insurance claim)
- What remedy options are available (e.g., referral, appeal, or request for review of the AIDA Entity's decision by the data subject)?

In addition, the document also calls for internal transparency from the perspectives of (1) ensuring the accuracy of explanations to data subjects and (2) winning the trust of internal stakeholders as a precondition for using AIDA. The following methods to ensure internal transparency are introduced (see Section 2.3).

- Explanation according to the learning method used by the model: method-specific (e.g., the method used to examine the weights learned in the layers of a neural network) or method-agnostic explanation
- Scope of explanation: depends on the quantity being explained (individual model prediction/overall aggregate drivers of model behaviour/internal component of the model etc.) internal component of the model) etc.)
- Explanatory method: the results must be summarized with respect to the

components of the model (the traditional method is to explain the local or overall characteristics of the model by explaining the impact of each feature on the overall model decision. For more complex models, the model internals such as learned weights and layers may be the subject of explanation, such as the branching structure of the decision tree or the visualization of the middle layer of a neural network. In addition, methods are also introduced to explain model outputs in terms of points from the underlying training data points that have contributed specifically to particular aspects of the learned model behavior, etc.).

The document also introduces the methodology for ensuring external and internal transparency (see Section 3).

6.5 Model AI Governance Framework for Generative AI

6.5.1 summary

The Information and Communications Media Development Agency (IMDA) and the AI Verify Foundation published the Model AI Governance Framework for Generation A I in May 2024.

Based on the risks of conventional AI, which analyzes and predicts insights from given data, and the new risks associated with generative AI, the purpose of this report is to show the actions required of policy makers, industry, researchers, and other stakeholders, based on their respective positions, in order to respond to the risks posed by generative AI, and is intended to strike a balance between responding to risks related to generative AI and innovation.

The Model AI Governance Framework for Generative AI consists of the following nine items, with each item indicating its importance and the actions to be taken.

① accountability
② data
③ Reliable development and implementation
④ Accident Reporting
⑤ Test execution and assurance
⑥ security
⑦ Reliable Content
⑧ R & D with safety and integrity
⑨ AI for the Public Good

6.5.2 Reference to transparency

Among the nine items in the Model AI Governance Framework for Generative AI, transparency is specifically mentioned in ”trusted development and implementation,” ”trusted content,” etc.

(1) Reliable development and implementation

It is noted that this item does not provide the information that should be obtained to assure that the model is reliable. It is said that best practices in development, disclosure, and evaluation will be important in the future, and that significant transparency will be key. On the other hand, it is also pointed out that transparency should be balanced with business security, ownership of information, and prevention of misuse of the system by malicious actors.

1) ”Food Labels.”

Regarding transparency, the report specifically describes a specific method in ”Disclosure” using the analogy of ”Food Labels”. Specifically, the report suggests that transparency can be achieved through standardized information disclosure like ”Food Labels.

- usage data

An overview of the type of training data and how that data was processed prior to training

- Learning Infrastructure

An overview of the infrastructure used to train AI. An overview of environmental impacts where possible (awareness of the issue of accelerating carbon emissions).

- Assessment Results

Overview and main results of the evaluation (comprehensiveness of evaluation methodology, subject matter, sufficiency consistency of tools, etc. are issues)

- safety measure

Technologies to correct biases, measures to prevent the leakage of sensitive information, and security measures that have been introduced.

- Risks and Responses

Risks identified for the model and actions taken to address them

- Intended use of the model

Documentation specifying the intended use of the model in question

- User Data Protection

Overview of how user data is used and protected

2) Implications for Industry and Government

The report suggests that industry take a more in-depth approach by building consensus on the fundamentals of transparency and considering certain standards. As an alternative to such consensus building, the report also suggests that there is room for policy makers to establish standards.

In addition, there is room for policymakers to define risk thresholds for risky areas related to matters of national security or high social impact to enhance transparency to the government and to help provide additional oversight.

(2) Reliable Content

In this item, he pointed out that it is difficult for users to distinguish between AI-made content and non-AI-made content, given the fact that generative AI can quickly mass-produce synthetic content. He raised issues such as the possibility of social threats, using the impact of deep fakes in elections as an example. In this regard, the report points out that transparency regarding the source of such content should be ensured so that users can use online content with appropriate information.

1) Means of Responding to Risk

The report states that the government and businesses are searching for technologies to address the above risks, and it cites Digital Watermarking and Cryptographic as examples. The report also suggests that technological solutions alone may not be sufficient, and points out the need for a mechanism with a certain level of enforcement by the government.

2) Issues Related to Policy Making

On the other hand, it is also pointed out that careful consideration is needed for the design of the system, taking into account that it is not realistic to target all content, that information on the source can be fragmented from the content, that consumers do not have a high level of understanding of the tool, and that the tool can be misused in the form of false certification, etc. It is also pointed out that careful consideration is needed in light of the fact that information on content can be fragmented, that consumers do not have a high level of understanding of the tool, and that the tool can be abused in the form of false authentication, etc.

3) Suggestions based on issues

In this regard, the report presents the roles expected of each party, noting the importance of cooperation with related parties based on the content lifecycle. In addition to suggesting the standardization of information related to the compilation of content, it also suggests that end users should have a deeper understanding of the source of information and that publishers and other business operators should take measures such as embedding identification symbols in content, displaying details of the source, preventing the misuse of the approval system, and narrowing down information so that users can understand it. The report also suggests the following measures to be taken by issuers and other business operators

6.6 Advisory Guidelines on the use of Personal Data in AI Recommendation and Decision Systems

6.6.1 Objective.

On May 1, 2024, the Singapore Personal Data Protection Commission (PDPC) issued Advisory Guidelines on the Use of Personal Data in AI Recommendation and Decision Systems (Guidelines on the Use of Personal Data in AI Recommendation and Decision Systems”). The purpose of the Guidelines is to provide guidance on the use of personal data in AI recommendation and decision systems. The purpose of the Guidelines is to ensure that systems incorporating machine learning models (hereinafter referred to as ”AI systems”) do not normally make autonomous decisions. The purpose of the Guidelines is to provide organizations with certainty as to when personal data can be used in developing and deploying AI systems, and to provide consumers with assurance as to the use of their personal data in AI systems, given that systems incorporating machine learning models (”AI systems”) are typically used to make autonomous decisions or to assist human decision makers through recommendations and predictions. providing assurances about the use of the data (Art. 1.2).

> 1.2 The purpose of the Advisory Guidelines on the Use of Personal Data in AI Recommendation and Decision Systems (”Guidelines ”) is to provide organisations with certainty on when they can use personal data to develop and deploy systems that embed machine learning models (”AI Systems”), and give consumers assurance on the use of their personal data in AI Systems. models (”AI Systems”), and give consumers assurance on the use of their personal data in AI Systems, since they are typically used to make autonomous decisions or assist a human decision-maker through recommendations and predictions.

The Guidelines also recommend that, in the interest of transparency, relevant information be provided at the time of data collection so that consumers can give meaningful consent to the collection of their personal data, and that the organization include in its policies the safeguards and practices it has in place to ensure the trustworthiness of its AI systems, especially when the results have a significant impact on consumers (Article 1.5). It recommends that the safeguards and practices that are in place be included in the organization's policies with respect to the organization (Art. 1.5).

> 1.5 To assure consumers that their personal data is being used appropriately, the Guidelines encourage organisations to be more transparent. To this end, organisations are encouraged to provide relevant information at the point of data collection so that consumers can give meaningful consent. They are also encouraged to include in their written policies about safeguards and practices they put in place to ensure that AI Systems are trustworthy, especially where the outcome has a high impact on trustworthiness. They are also encouraged to include in their written policies about safeguards and practices they put in place to ensure that AI Systems are trustworthy, especially where the outcome has high impact on consumers.

6.6.2 Scope of Application of Guidelines for the Use of Personal Data

The Personal Data Protection Act ("PDPA") is a broad law that applies to all collection and use of personal data by organizations, as well as to the collection and processing of personal data for the development, testing, and monitoring of AI systems or as part of their deployment process .... The Guidelines on the Use of Personal Data for AI Recommendation and Decision Making Systems apply to situations where the design and deployment of AI systems involves the use of personal data in scenarios covered by the PDPA.

6.6.3 Ensure transparency in the development and use of AI systems

The Guidelines on the Use of Personal Data in AI Recommendation and Decision-Making Systems indicate that during the development of an AI system, personal data can be used without the consent of the individual and shared with third parties by applying exceptions (Business Improvement Exceptions and Research Exceptions). While indicating that personal data can be shared with third parties (Art. 5.6), they also require that the consent of individuals be obtained for the collection and use of personal data by AI systems to provide recommendations, predictions, and decision-making (Art. 9.1). And in seeking this consent, the user must be notified in advance of the purpose and intended use for which the personal data will be collected, in accordance with Article 20 of the PDPA (Art. 9.2). Among other things, in order for individuals to provide meaningful consent, such notification is required to provide information on the type of personal data to be collected and processed and the purpose of the processing (e.g., film recommendations), and the Guidelines further recommend that the following information be provided to the extent practicable (Art. 9. (Articles 3-9.5).

> a) The function of their product that requires collection and processing of personal data (e.g., recommendation of movies); and
>
> Product features that require the collection and processing of personal data (e.g., movie recommendations)
>
> b) A general description of types of personal data that will be collected and processed (e.g., movie viewing history); and
>
> General description of the types of personal data collected and processed (e.g., movie viewing history)
>
> c) Explain how the processing of personal data collected is relevant to the product feature (e.g., analysis of users' viewing history to make movie recommendations); and make movie recommendations); and
>
> Explanation of how the processing of collected personal data relates to product functionality (e.g., movie recommendations based on analysis of user viewing history)
>
> d) Identify specific features of personal data that are more likely to influence the product feature (e.g., whether movie was viewed completely, viewed multiple times, etc.). multiple times, etc.). d) Identify specific features of personal data that are more likely to influence the product feature (e.g., whether movie was viewed completely, viewed multiple times, etc.).

In addition, the Guidelines define Accountability Obligation as the manner in which an organization fulfills its responsibility for the personal data it collects or obtains for processing, or for the personal data under its control, and the actions an organization must take to fulfill this obligation are set forth in Articles 11 and 12 of the PDPA (Article 10.1). The actions to be taken by organizations to fulfill this obligation are set forth in Articles 11 and 12 of the PDPA (Art. 10.1). And where Article 12 of the PDPA requires organizations to develop policies and practices to fulfill their obligations under the PDPA, it states that organizations using AI systems should include in their written policies relevant practices and safeguards to be transparent and achieve fairness and reasonableness (Art. 10.3, ibid.). Furthermore, while Article 12(d) of the PDPA requires organizations to provide information about such policies and practices to individuals upon request, the reason for the existence of such external communications with consumers is to build trust with data subjects by demonstrating accountability in complying with the PDPA Because of this, organizations should consider disclosing such written policies in advance through their websites, rather than only upon request (Art. 10.4).

7. summary

We have looked at the rules in transparency (information disclosure) in AI in various countries above. We would like to point out three points in this context.

The first point is the need to discuss the specifics of transparency. The definition of transparency itself is not the same in each country. Although the concept is generally centered on the disclosure of information, a look at the rules in each country regarding what information is to be disclosed, to whom, and for what reasons shows that there is also a gap in the definition of transparency. In comparing the rules of different countries, it is necessary to consider not only whether the concept of transparency exists in the rules, but also the content and scope of information that is required to be disclosed based on transparency. The following is a list of information that is required to be disclosed under the representative guidelines and rules of each country for reference. It is clear that there is a divergence in the content of information required to be disclosed based on transparency in each country. Such differences in the information required for disclosure may be due to differences in the roles expected of transparency and what is desired to be achieved through transparency in each country.

The second point is who or what should be the subject of disclosure: In EU AI law, the main rule is disclosure from the developer to the deployer, while in Japan and Singapore, the main rule is disclosure to users and the general public. In the UK, there is a standard for information disclosure by public authorities (Algorithmic Transparency Recording Standard), and rules for disclosure mainly by public authorities are in place. These differences indicate that there can be differences among countries in prioritizing the values they wish to realize through transparency. In other words, the EU may place more emphasis on the implementation of user protection measures by downstream deployers, Japan and Singapore may place more emphasis on measures taken by users themselves, and the UK may place more emphasis on the sharing of best practices by public authorities, where competition is less of an element. The origin of these differences must also be examined.

The third point is that it seems necessary to consider on what grounds certain information should be required to be disclosed. However, the accuracy of AI varies depending on the environment in which it is used. It may be necessary to examine the meaning of such disclosed information and reconsider the information that really needs to be disclosed.

The perspectives described above are not limited to transparency. While comparisons of rules in various countries have been actively made recently, we hope that deeper analysis from the perspectives mentioned above will lead to more appropriate AI-related rulemaking.

International rule comparison on AI transparency

July 2024

AI Law Study Group
International Exchange Subcommittee