AI Algorithm Transparency Toolkit
A Proposal for a Governance System to Enable Society to Accept and Benefit from AI-based Innovations

YUKI Tonfi[*†], OKAMOTO Masayuki[‡], MURAYAMA Hiraku, YAJIMA Kirihito[§], NISHIGAKI Yuta[** ††]

Summary

To optimize the benefits of AI innovation, it is crucial to establish and manage a flexible technical, organizational, and social framework that allows stakeholders to acknowledge AI-related risks and adapt their interests accordingly—an agile governance system for AI. This paper addresses the complex and contentious issues surrounding AI governance, focusing on the organization of overarching and systematic approaches for the transparency of AI algorithms, and proposes a practical toolkit for implementation.

In the absence of a systematic organization, regulations and norms are formulated for various problems arising from innovation, leading to a continuous addition of rules without a clear blueprint. This approach lacks predictability and reduces "transparency" to a mere formality, deviating from its original purpose. To ensure that businesses and government agencies can effectively apply these regulations, it is imperative to pursue a systematic approach founded on a unified concept, generality, clarity, and flexibility, allowing for discretion in decision-making.

This toolkit comprises a comprehensive set of disclosure items and case studies, taking into consideration discussions among international regulators, a wide array of risk events generated by various AI algorithms, and prior research on AI algorithms. It also embraces the perspective that the level of transparency achieved by businesses and government agencies can positively impact not only their societal credibility but also other metrics, such as satisfaction for the product that the AI algorithms are incorporated.

Designed to be highly user-friendly for both businesses and government agencies engaging in self- and co-regulation, the toolkit presents disclosure items in a list format, allowing for discretion and flexibility when considering AI algorithm providers, users, and risks. Additionally, it includes practical use cases to further enhance its applicability.

Table of Contents

[**] SmartNews, Inc. / ZeLo, a Foreign Law Joint Enterprise
[†] (Main email address) tonghwi.soh@gmail.com
[‡] Toyota Motor Corporation
[§] Aflac Life Insurance Japan Ltd. / Aflac Digital Services Company Ltd.
[**] Graduate Schools for Law and Politics, Faculty of Law, The University of Tokyo
[††] Program assistant

1 Why Transparency of AI Algorithms is Required

1.1 Introduction

This paper proposes a solution from the perspective of *transparency* of Artificial Intelligence (AI) algorithms in order for society to control the risks and adverse effects of AI that are becoming apparent and to maximize the benefits of AI in this age when AI has naturally penetrated society as a result of the third AI boom. Innovation often disrupts existing systems, structures, value systems, business models, etc., in some way, and there is no end to the misfortune caused by innovation being stifled as a result of society's excessive fear of the risks involved. A report published by the Ministry of Economy, Trade and Industry (METI) [1] in 2021 describes the need for agile governance of innovation as "the design and operation of technological, organizational, and social systems by stakeholders with the aim of maximizing positive impacts while managing risks arising in society at a level acceptable to stakeholders." In order to balance AI-based innovation and the risks it entails, risks need to be appropriately distributed among stakeholders at an acceptable level. As an example of such governance, the perspective of "transparency" is required. Transparency will help maximize the impact of innovation by controlling AI risks to an acceptable level and building trust throughout society through appropriate communication with society and users.

However, although the topic of transparency of AI algorithms has already been discussed internationally, there are a number of different proposals. While ensuring transparency is important, as we will see in this paper, if the degree of transparency is wrong, the act of disclosure itself may even reduce users' trust and satisfaction. In addition, regulations must not be added without systematic discussion each time various problems caused by innovation are addressed. Otherwise, the means of transparency to build governance and trust and protect the right to self-determination will become an end in itself, and as a result, responding to low-strength, low-quality disclosure items that have been repeatedly added without blueprints will become the supreme objective. As self-regulation and co-regulation are becoming more common as a way of regulation that can withstand the speed of innovation, a method of achieving transparency that is easy for operators to understand and realistic for them to take voluntary action must be proposed.

This paper addresses these issues by attempting to systematically organize the discussions to date and proposing a toolkit version 1.0 that can withstand practical use. Through this toolkit, we propose a systematic and convenient collection of disclosures and case studies based on various risk events posed by AI and prior research on AI, while also addressing the latest regulatory discussions in various countries. This paper stands as a guide or handbook for utilizing the toolkit.

Throughout this paper, AI refers to a generic term for software and systems that "have the ability to change their own output and programs through the process of utilization by learning data, information, and knowledge, etc." [2] Although this toolkit mainly focuses on AI algorithms using machine learning and deep learning, the discussion points are generally the same for other AI algorithms such as rule-based and knowledge-based algorithms. In addition, the term "AI algorithm" in this paper refers to the process by which AI solves a problem, i.e., "a finite number of procedures by which AI solves a

[1] Ministry of Economy, Trade and Industry, "Government Innovation Ver. 2: Toward the Design and Implementation of Agile Governance," p. 95 https://www.meti.go.jp/press/2021/07/20210730005/20210730005-1.pdf (2023) (viewed January 28, 2023)

[2] Ministry of Internal Affairs and Communications, AI Network Society Promotion Council, "AI Utilization Guidelines - Practical Reference for AI Utilization," p. 4 (https://www.soumu.go.jp/main_content/000637097.pdf (viewed January 28, 2023)).

mathematical problem (a problem whose solution is fixed and computable)."[3]

1.2 Risks and adverse effects of AI-based innovation

To identify the issues to be addressed in this paper, we would like to look at some examples of the risks or harms of AI innovations.

It has been noted that one social media A, has developed algorithms that maximize user engagement metrics, which may be detrimental to the mental health of its users, especially the younger generation. For another social media B, it is noted that the ranking of posts in feeds may be artificially manipulated, in some cases with some political agenda, in ways that are beyond the control of users and society.

Human resource service C, which provides job matching services, has been found to give preferential treatment to job seekers of a particular race or gender. Image recognition service D encountered a problem in which a black person was judged to be a gorilla.

In the restaurant rating information service E, a change in the rating algorithm could cause significant harm to restaurants. The same is true for the impact of changes in the algorithm of search engine F on media businesses and viewers.

Image-generating AI service G can produce countless similar works against the will of a particular artist, and another video-generating AI service H has successfully interfered with elections by producing fake videos of politicians.

1.3 Transparency: most frequently mentioned in AI principles and guidelines

Discussions have already begun on how human society should tackle the above-mentioned issues. According to a report by the Ministry of Internal Affairs and Communications[4], "transparency and accountability" is listed in 35 (about 90%) of the 40 AI-related development and utilization guidelines published by governments and researchers around the world. While some countries and organizations did not mention "privacy" and "fairness," which are other important issues, "transparency and accountability" is the only item mentioned in all the guidelines by all countries and organizations.

---

[3] "Algorithm." Merriam-Webster.com Dictionary, Merriam-Webster, https://www.merriam-webster.com/dictionary/ algorithm. Accessed 13 Jan. 2023.

[4] Ministry of Internal Affairs and Communications, AI Network Society Promotion Council, "Report 2022 - Further Promotion of 'Safe, Secure and Reliable Social Implementation of AI'" (July 25, 2022), p. 78 https://www.soumu.go.jp/main_content/000826564.pdf (2023, Jan. 28, viewed on January 28).

	# of Guidelines	Human Centric	Human Dignity	Diversity and Inclusion	Sustainable Society	International Cooperation	Appropriate Use	Education and Literacy	Intervention of human judgment Controllability	Proper learning (training data quality)	Collaboration among AIs	Safety	Security	Privacy	Fairness	Transparency, Explainability	Accountability	Robustness	Responsibility	Traceability	Monitoring, Audit	Governance	Miscellaneous
USA	6	1	2			1	1		1	1		3	3	1	4	5	4	1	2	2	3		
Canada	1							1								1							1
UK	6		2		1			1		2		1	1	2	5	5	5	1	2		2		1
Italy	1									1				1		1							
Netherland	2		1				1		1	2				1	2	2	1		1	1			
Sweden	2		1	1			1	1					1	1	1	2							
Denmark	2		2	2			1		1			1	1	1	2	2	1	1			1		
Germany	3	1	2	2	3				1	2		2	2	2	2	3	2	3	1	1	1		
Norway	2		2		1		2		1			1		1	2	1	1	1		1	1		
Finland	2		2	1			1		1			1		1	1	2	1		2	2	1		
France	1								1						1	1	1				1		
India	1		1	1								1	1	1	1	1	1		1				
Korea	1		1	1		1						1		1	1	1	1	1					
Singapore	2		1												2	2	1						
Australia	3		3	2	2	2	2	2	3	1		3	2	2	3	2	2	2		2	3	3	
EU	1	1	1									1	1	1	1	1	1		1		1		
Intl. Org.	2		2	1	2	1	2		1			2	1	2	2	1	2	2	2	1	1		
Total	2		2	2	2	1	2	1	1		1	2	1	2	1	2	1		1	1	1		
	40	3	25	13	11	6	13	6	12	9	1	19	14	20	31	35	25	12	13	11	16	3	2

Table 1 AI Network Society Promotion Council, Ministry of Internal Affairs and Communications, "(2) Verification based on comparison with principles, guidelines, etc. overseas" (cited from "Report 2022: Further Promotion of 'Safe, Secure, and Reliable Social Implementation of AI'", p. 78)

Beyond these guidelines, the EU has already initiated regulations to achieve some AI transparency with the General Data Protection Regulation (GDPR)[5] and is working on further regulations such as the AI Act[6], the Digital Services Act (DSA), and the Digital Market Act (DMA)[7]. Japan has also introduced co-regulation in terms of transparency in the Act on Improving Transparency and Fairness of Digital Platforms (DPF Transparency Act). Already, transparency of AI algorithms is not an ideal that should be realized someday but a demand from society that should be addressed in reality.

1.4 Possibility of resolving issues through transparency

Why is transparency in AI algorithms being called for so loudly?

In order to create a society in which humans are in harmony with AI, humans need to understand AI. Innovation, not limited to AI, comes with risks, and those risks must be evaluated and controlled so that innovation can permeate society. For humans to coexist with the risks of AI, it is essential to foster trust in AI. Human society must be able to understand that AI is safe and secure if handled properly. For example, food products we eat are labeled with information such as nutritional content, ingredients, country of origin, and expiration date. This transparency is necessary to ensure the safety and reliability of the food products we purchase. More theoretically speaking, it is a measure to protect the right to self-determination. The examples of risks and adverse effects of AI-based innovation described at the beginning of this paper can find solutions if businesses consciously communicate with society about the risks posed by AI and how to control them. A discourse exists that AI is inherently value-neutral and more trustworthy than humans, but it misunderstands the nature of the problem. It has been

[5] European Commission "EU data protection rules" https://commission.europa.eu/law/law-topic/data-protection_en ((viewed January 28, 2023).

[6] European Commission "A European approach to artificial intelligence" https://digital-strategy.ec.europa.eu/en/policies/ european-approach-artificial-intelligence" (viewed January 28, 2023).

[7] European Commission "The Digital Services Act package" per DSA, DMA respectively https://digital-strategy.ec.europa.eu/en/ policies/digital-services-act-package (viewed January 28, 2023).

pointed out that it is humans who develop AI, and as a result, introduce "systemic biases"[8] and "reinforce historical discrimination."[9] The issue is not whether AI is more trustworthy than humans but rather the transparency of AI as a basis for individuals and society to judge whether AI developed by humans can be trusted.

However, transparency of AI algorithms does not prevent businesses from offering benefits to individual users and society. Rather, businesses that develop and provide AI services can expect to benefit from them too. Businesses can identify the factors that differentiate them from their competitors by ensuring the transparency of AI algorithms. Continuing to use the food analogy, this is similar to using organic and pesticide-free vegetables and direct delivery from the place of production as differentiating factors. Efforts to achieve transparency can also be linked to risk management for businesses. The toolkit proposed in this paper can serve as a checklist to examine what organizational or technical measures to take after systematically organizing the risks posed by AI. By determining how businesses can control the assessed risks and communicating with society, it will be possible to work with society to resolve any risks that become apparent. As a result, for businesses that develop and provide AI, it is equivalent to practicing the risk management process from an overarching perspective.

Businesses sometimes express the concern that disclosing their AI algorithms may lead to the risk of being hacked by users or of trade secrets becoming known to competitors. However, the transparency of AI algorithms proposed in this paper does not mean disclosing the code in detail. Such disclosure would not explain transparency to the majority of users, as only a limited number of researchers and developers would be able to understand what it means. This paper proposes, based on international discussions, a more abstract and logical explanation of features and their weighting. What is required is not the disclosure of a secret sauce recipe, but rather the disclosure of the ingredients, cooking method, and seasonings used for the dish.

2 Toolkit Overview

2.1 Purpose and role of the toolkit

As we saw in the previous section, although the need for transparency of AI algorithms has been advocated in various places, concrete measures have been left to voluntary efforts or preceded by specific regulations in individual countries, and practice has begun to progress without an overall systematic organization and understanding. The toolkit proposed in this paper (the "Toolkit") aims to serve as a catalog of responses with contents that can withstand practical application, based on a systematic organization of the transparency of AI algorithms, reflecting the latest discussions. In turn, the issues that each country's regulations and self-regulations by operators seek to resolve can be understood from a bird's-eye view by referring to this Toolkit.

The reason for proposing a toolkit rather than principles and guidelines for transparency is that the content and degree of transparency required will vary depending on the entities, purposes, methods, and situations in which AI is used. The aim is not a grand ideal, but a catalog that can be used by each stakeholder depending on the context. Therefore, this toolkit is a collection of options to ensure transparency and should not be taken as a strict disclosure item. Let us explain how to use the toolkit from the perspective of building trust and communication between the business that develops and provides AI and the users and society.

---

[8] Kashin, K., King, G., & Soneji, S. (2015). Systematic Bias and Nontransparency in US Social Security Administration Forecasts. Journal of Economic Perspectives, 2(29), 239-258 . (https://www.aeaweb.org/articles?id=10.1257/jep.29.2.239 (viewed January 28, 2023))

[9] Kroll, J. A. (2015). Accountable algorithms.(Doctoral Dissertation) Princeton University (https://dataspace.princeton.edu/handle/88435/dsp014b29b837r (viewed January 2023) (viewed 28 January 2023))

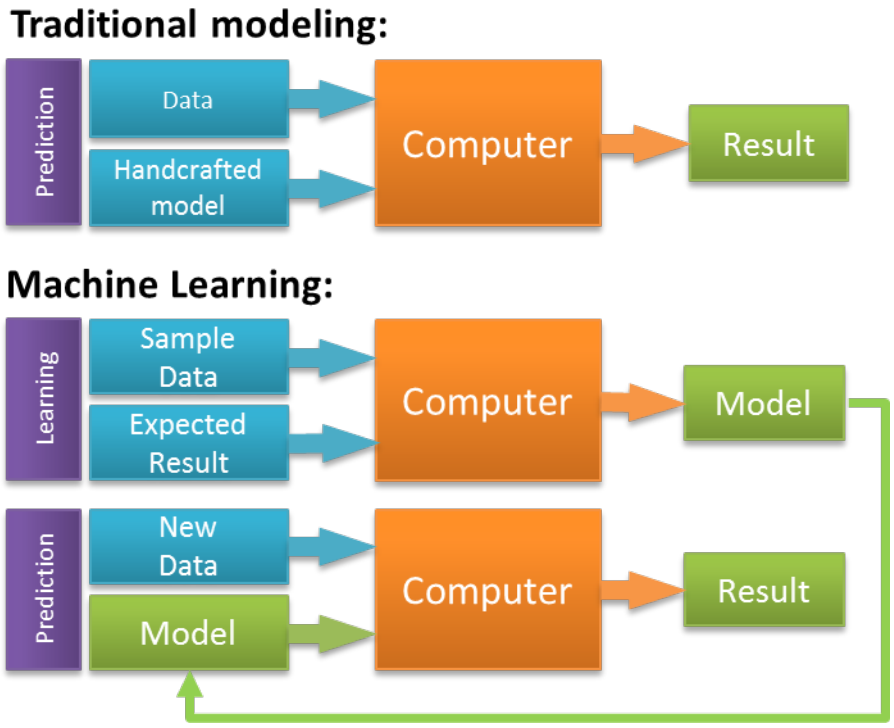
2.2 Scope of the toolkit

As mentioned in Section 1.1, this paper refers to AI in terms of software and systems using machine learning models and deep learning models. It does not restrict the purposes or objectives for which AI is used, so its contents can be widely used for all software and systems using AI.

However, there are some issues with AIs that cannot be solved with transparency. In particular, manufactured goods or products equipped with AI that involve risk to life or body, or services for which users have few alternatives (e.g., automated driving, digital platforms, etc.) must pursue the safety of users in the first place, and transparent communication about the existence of risk alone will not lead to practical implementation.[10] In this sense, this toolkit is structured on the premise that AI and risks go hand-in-hand.

2.3 Structure of the toolkit

As shown in Figure 1, algorithms of machine learning models are different from traditional programming, in which humans develop an algorithm by specifying commands and conditions, in that the model itself is generated by a machine based on data.[11] Since an understanding of this point is essential for the use of this toolkit, we would like to review the differences between the two specifically using Figure 1.



[10] Even if the "risk of this self-driving car getting into an accident" were disclosed, the only thing the user can do is not get in that car.

[11] We would like to clarify the definitions of "algorithm" and "model" in machine learning, just to be clear. A machine learning algorithm is a procedure that learns a processing method from input data or returns a processing result in response to input data, depending on the purpose, such as classification or prediction. Depending on the purpose, algorithms such as logistic regression, decision trees, and neural networks are used. A machine learning model is a program that outputs some kind of identification or judgment result for new input data, which is the result obtained from processing training data by a machine learning algorithm. Even if the same algorithm is used, the output model will be different if the data to be learned is different. The terms used in this text are also based on this definition.

Figure 1: Comparison of traditional programming and machine learning algorithms (adapted from ZEISS [1 2])

In traditional software development, such as rule-based programming, the output (the result of recognition and inference; "Result" in Figure 1) is obtained as the result of processing data by a program created by an expert on a computer ("Handcrafted model" in Figure 1).

In the case of training phase of machine learning algorithms, on the other hand, a model showing the input-output relationship ("Model" in Figure 1) is obtained from a sample of input data and expected output (correct answer) for that input (training data; "Sample Data" in Figure 1). A machine learning algorithm is used for obtaining the model from the data. This model is then used to obtain recognition or inference results ("Result" in Figure 1) for new input data. As a simple example, the curve $f(x)=w_0+w_1x+w_2x^2 \ldots + w_Mx^M$ (w_0, ..., w_M are coefficients, x^n is nth power of x), the method of estimating coefficients from a given combination of (x,f(x)) (e.g., least squares method) is the trained algorithm, and the function to which specific coefficients are fitted is the model of the learning result.

Furthermore, human involvement continues to be significant in that the entity developing the model checks the validity and quality of the output (the results extracted by the model), develops additional models, and repeats this process. In Seaver's words, "an algorithm is not one small independent box, but a large network of hundreds of people who tune the system, replace parts, and experiment with new arrangements." [1 3]

Therefore, simply disclosing the code of the model or explaining only about the model does not solve any problems.

This toolkit views AI algorithms as being formed by the organic coordination of the five elements of data, models, outputs, human involvement, and overall system design. As shown in Appendix 1, we attempt to ensure the transparency of the entire AI based on these perspectives. The details of each item are described in the following sections.

3 Toolkit Description

In discussing the transparency of AI algorithms, we would first like to outline this development and delivery process based on Figure 2.

[1 2] ZEISS, "The Relation between Computer Vision and Machine Learning," https://blogs.zeiss.com/digital/page/2/ (viewed January 28, 2023).

[1 3] Seaver, N. (2019). Knowing algorithms. DigitalSTS, 412-422. https://digitalsts.net/wp-content/uploads/2019/03/26_Knowing-Algorithms.pdf (viewed January 28, 2023)

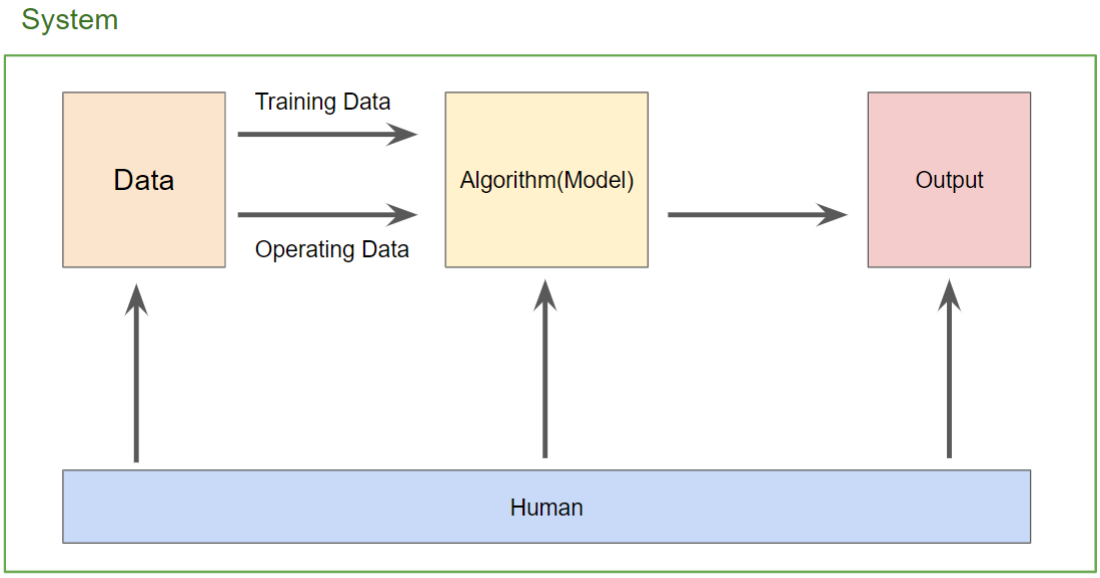


Figure 2: Overview of the AI development process (prepared by the authors)

First, training data (Training Data) is needed to develop AI models. Based on the training data, the AI constructs a model, and then checks whether the model can withstand actual operation using test data. Then, in actual operation, data (Operating Data) is input to the model as actual input, and the model produces an output (Output) based on the input. All of these processes involve human involvement (Human). The concept and design that crosses each of these elements is referred to as the overall system design (System) in this toolkit.

Appendix 1, "Transparency Disclosure Items List for AI Algorithms," lists the disclosure items in ascending order based on the broad life cycle of AI development, and specifies which of the above five elements each item is related to. The following section looks specifically at the transparency that should be realized for each of the five elements. For convenience of explanation, we will look at data, models, outputs, overall system design, and human involvement, in that order, but this order is basically random.

3.1 Data

3.1.1 Overview of training data

In developing an AI system, an overview of the training data used (including data used for testing. Same below) is disclosed. For example, for a text generation AI, the training data used may vary depending on the large-scale language model used. For image generation AI, the characteristics of the image dataset used should be described in a form that is easily understood by users and the general public.

It is also important to devise different disclosures for different learning methods. For example, in the case of supervised learning, what kind of data is used as training data and what is learned? In the case of unsupervised learning, what is the model expected to do? In the case of reinforcement learning, it may be necessary to clarify the incentives to be given to the model.

Even if it is not possible to disclose all variables and items that build the data set, it may be necessary to clearly state if there are any special or sensitive items involved.

3.1.2 Sources of training data

Provide access links to datasets that are publicly available in a form accessible to users and researchers.

If the data is open data, this should not be too difficult, but if not, a sampled dataset that is consistent in content with the overall dataset could be provided.

3.1.3 How the training data was obtained

Describe how the learning data was obtained.

In most cases, the personal information protection regulations of each country already require disclosure of the acquisition method and purpose of use of personal information, at least at the acquisition stage. This disclosure should also be required at the AI use stage. In addition, there may be cases where data other than personal information (e.g., statistical data, text data) should be disclosed as well.

3.1.4 DEIB (diversity, equity, inclusion, and belonging) policy for training data

Describe the policy and views of the business regarding diversity, equity, comprehensiveness, representativeness (is the AI appropriately representative of the population for which it is intended to be used), and inclusiveness that the training data takes into account in developing the model for the AI.

When an AI has bias, one of the factors is that the training data itself is biased. Therefore, it is important to explain how bias is eliminated so that AI can withstand actual operation, or if bias remains, to explain the reason for it. This will not only contribute to the convenience of users, but also enhance the awareness of service providers during development.

For example, when providing AI for the Japanese market, a dataset that is clearly biased for the U.S. society may be acceptable from the perspective of diversity and inclusiveness in Japanese society. In such a case, it is important for the provider to clarify the results of its consideration of sufficiency from the DEIB perspective in the relevant market, rather than pursuing the use of a diverse data set that would be acceptable to any society.

In addition, explanations regarding diversity, comprehensiveness, and representativeness may be possible by explaining data that were intentionally excluded from the training data.

3.1.5 Operational data

Describe an overview of the data that will be used when AI is actually utilized.

If all of the items related to training data described above can be applied to operational data as well, they can be organized into one large item called "data". On the other hand, if, unlike training data, data directly input by the user, for example, is used as operational data, a separate description of the data actually used may be required from training data.

3.2 Model

3.2.1 Phases of use

Explain in which phases the AI models is used in the development of products and services. For example, by clearly indicating the phases such as ideation phase, research phase, beta phase, or production phase, users will be able to adjust their expectations and usage of the AI.

3.2.2 Learning method

Describe the type of learning method used to develop the model. The description and granularity of these explanations – on topics such as machine learning, deep learning, expert systems, rule-based, etc. - should be customized on a case-by-case basis, adjusting to the understanding level of the users and researchers.

3.2.3 Explainability

AI creators should describe their understanding of the model to the extent that they can.

For example, due to the black-box nature of deep learning models, there could be cases where verbalizing abstract parameters and features is not entirely possible. However, considering that the purpose of transparency is to build trust in AI products and to protect their rights, including their right to self-determination, there are many issues that can be resolved by ensuring transparency by providing some level of explanation at a granularity level that is understandable by humans." Explainable AI (XAI), which has been introduced in recent years, is a notable example of such efforts. It is safe to assume that such resources can help offload

excessive accountability from operators and offer alternative ways to explain AI models on the behalf of people.

3.2.4 Parameters/Features and their purpose/reason

Describe the parameters/features that make up the corresponding model. It is preferable to explain the operator's understanding of their purpose and reason for using a certain model. A parameter or feature is a quantitative expression of the characteristics of a data set. This paper does not seek to strictly define them, but rather to define them in terms of the quantifiable characteristics and settings that AI algorithms take into account when generating certain output results. The quantification of which parameters are important is called a weighted parameter (to be explained later). In AI product development, the main issues are how to design the parameters or obtain them from data, and how to set or optimize weights. A popular example used is an AI algorithm that is used to predict ice cream sales. An AI model is built by learning the relationship between sales and various factors such as weather, temperature, number of products, and location (e.g. distance from the nearest station). Each of these factors make up a parameter. The parameters of the model will be determined by the number of products and sales data. As is often seen in deep learning, there may be cases where effective parameters are obtained as the model learns through the data, in addition to people manually designing individual parameters.

However, it is not required to disclose all detailed parameters and other information. Disclosure of detailed parameters may make it difficult to understand for users and lead to unintended disclosure of a business' "secret sauce". Rather, it is most practical to verbalize the important parameters considered by the model, which are abstracted at a level that is understandable to users, together with the weightings described in the next section. Going back to the previous example, users do not require an explanation of "the weather classified into 18 types, based on the following technical definitions." Instead, they can understand what is necessary via a simple "today's weather" parameter. When the AI is learning parameters using deep learning, it is possible to verbalize and disclose only the learning methodology. In addition, by explicitly explaining factors that do not affect the model but could easily be misunderstood as having an impact on the model cany help facilitate appropriate usage of AI algorithms.

Perhaps the most informative regulation in this regard as of January 28th, 2023, is the EU Regulation on platform-to-business relations (P2B Regulation)[14]. The Regulation requires online platform operators to achieve transparency of key algorithms and parameters in order to achieve fairness and transparency in transactions between operators, including SMEs (small and medium-sized enterprises) on online platforms. Specifically, the Regulation requires that key parameters which affect rankings be determined and explained. If there are many parameters in the AI algorithm, the practice is to classify them into several categories, identifying major categories that play a decisive role. It is also considered best practice to describe the decision-making process itself leading up to the determination of the major parameters. In addition, a list of commonly used parameters is published[15], making it easy to understand what are potential "parameter" descriptions. For example, "quality of content (reciprocal links to websites, richness, quality and number of languages supported)," "geographic

[14] European Commission "Platform-to-business trading practices" https://digital-strategy.ec.europa.eu/en/policies/platform- business-trading-practices" (viewed January 28, 2023).

[15] EUR-Lex "Guidelines on ranking transparency pursuant to Regulation (EU) 2019/1150 of the European Parliament and of the Council "(Annex 1) https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020XC1208%2801%29 (viewed 28 January 2023).

proximity," "popularity of products and services," and "availability of inventory" are listed.

Although the Digital Platform Transparency Act in Japan also requires disclosure of key parameters, the P2B Regulations are unique in that they further require an explanation of the reasons for disclosure. Depending on the type of AI algorithm, if the purpose, reason, and intent of the AI algorithm can be disclosed, users and society will be able to understand how to utilize the AI algorithm in a deeper manner. The following are some examples of actual disclosures.

For example, Google LLC, on its official website[16], lists "user relevance," "quality of app experience," "editor rating," "advertising," and "user experience" as key factors to explain the app rankings in Google Play. Having said that, Google notes that "These main factors impacting ranking are weighted differently based on where on Google Play a user is looking, the device they are on, and their personal preferences." In order to promote a deeper understanding of each item, the report goes on to explain in detail why and how such factors were chosen as "key factors" in the study.

For merchants, Amazon Japan G.K.[17] discloses that the main parameters that determine the product ranking in its Amazon Listing Services online mall are factors such as the match rate of text and other information with the product information, price, points, stock availability, product line, and sales history". For general users, it discloses the following: "customer behavior (e.g. frequency of product purchases), product information (e.g. product name, price, product description), availability, delivery time, various fees (e.g. delivery fees), and whether the product is of potential interest (e.g. a new product). The site is unique in that it places different explanations for merchants and for users.

On its official website[18], Yahoo Japan Corporation mainly considers products with a speedy reliable shipping icon in the mechanism for determining search rankings and what is displayed in particular slots. Furthermore, as "other factors to be considered," it states that "the relevance of search terms, the number of purchases, the number of customers who purchased the product, the number of units sold, the number of product reviews, the number of store ratings, the number of store ratings, and the number of store customers who purchased the product are all considered. At the same time, the report addresses potential misunderstandings by clearly stating that "payment of sales promotion fees, etc. from store owners does not directly affect the ranking position." In addition, the report[19] carefully explains the reasons and perspectives that led to the extraction of these key parameters. Namely, "we extracted the major factors that determine the display order of our online mall's 'Recommended Order' from approximately 120 factors that have a significant degree of influence on the order. This is based on the policy that factors that can be used as a reference by store owners who wish to improve their display order should be made public. We have extracted the factors that have the greatest impact on ranking from among approximately 120 factors. We have made the items as consistent as possible with the numbers available in the store tools (number of

---

[16] Google LLC, "App Detection and Ranking," https://support.google.com/googleplay/android-developer/answer/9958766?hl=ja (viewed January 28, 2023)

[17] Amazon Japan G.K., "Summary of Periodic Reports by Specific Digital Platform Providers for Fiscal Year 2021," pp. 50, 66 https://www.meti.go.jp/press/2022/12/20221222005/20221222005-c.pdf (January 28, 2023) (Viewed)

[18] Yahoo Japan Corporation, "Introduction to Transparency Initiatives," https://business-ec.yahoo.co.jp/shopping/digitalplatformer/ (viewed January 28, 2023)

[19] Yahoo Japan Corporation, "Summary of Periodic Reports by Specific Digital Platform Providers for FY2021," p. 43.

customers, number of orders, number of reviews) and the scores disclosed as "store performance" (number of orders, average store rating score, etc.), so that store operators can easily improve their store." Furthermore, as an explanation based on the characteristics of the machine learning algorithm, "since the content and weighting of the factors that determine display rankings change daily through machine learning, it is almost impossible to accurately disclose all of them at all times, but when we add a new major factor that has a large impact on rankings, we always update the disclosure.

3.2.5 Weights and their purpose and rationale

This section describes the weights or coefficients for each parameter. Weights are coefficients that quantify which parameters are important. In the previous example of ice cream sales forecasting, if the emphasis is on temperature, the weight for the parameter representing temperature will be increased, and if the emphasis is on location, the corresponding weight will be increased. The weights may be set by an expert on the problem that needs to be solved, or they may be adjusted and optimized by learning from the data. As with the parameters and features, the weights should describe the operator's understanding of the purpose and reason for the weights.

Parameters are not often dealt equal to each other in the model, and an explanation may be required that takes into account their interrelationships and relative importance, so that users can understand the differences in importance between parameters. However, since weightings can change on a daily basis and detailed adjustments are often kept secret within a business, a description of relative relationships is often preferred over individual existing descriptions.

The EU P2B regulation, for example, does not require the disclosure of precise weighting indices or coefficients between parameters. Since what is beneficial to users is being able to assume how parameters should behave, limiting the disclosure to relatively important parameters would be sufficient to achieve this purpose. Therefore, instead of describing this item independently, a possible method is to extract parameters that have a large impact based on relative comparisons among parameters as major parameters or important elements, together with the parameters and characteristics in the previous section. Parameters are important because they affect the output extracted by the model (e.g. display rankings, rating results, screening results, judgment results, etc.), the matters that are particularly important to the operator (through the model, the operator's thoughts and philosophy on the AI algorithm are reflected), or one particular measure would be the factor that should be considered by the user for optimal use of the AI algorithm (through the user's optimization behavior, the operator's ideology/philosophy to the AI algorithm is reflected).

Some of the disclosure cases already mentioned in the previous section have basically adopted the policy of explaining key parameters, which seems to be generalized in the future. On the other hand, for AI algorithms used by public organizations or AI algorithms that are considered to have high risks, not only the main parameters but also the relationships among them should be described, and in such cases, separate and independent explanations should be provided.

3.2.6 Different treatment for different user categories, including billing

This section explains how different treatment may be given to different categories of users, in cases where parameters and weighting may vary depending on the billing or other actions or attributes of the user. For example, there may be cases in which different treatment is applied to categories based on whether or not the user is charged or different categories based on service provider, age, gender, total usage period/frequency, and so on. This also includes explaining when a service operated by a service provider (including its group companies) and other services are treated differently in the AI algorithm.

In fact, the EU P2B Regulation has a regulation that requires disclosure of how the direct or indirect payment affects the ranking algorithm. Similarly, regulations regarding in-house preferences are also included in the P2B Regulation and the DMA.

For example, let us assume that a platform operator uses AI to adjust its recommendation engine and ratings. A social media platform may disclose that it prioritizes the posts of charged users in terms of display order and that it makes it easier for replies and shared posts to be displayed. There are also operators that have policies that do not use certain AI algorithms for users of a certain age.

3.2.7 Proper Procedures to Make Changes

This describes the process to be followed by service providers when AI algorithms are changed.

If changing the AI algorithm substantially affects users, e.g., affects ratings or changes the results of recommendation engines, prior notice, purpose of the change, and expected impact of the change should be described from the viewpoint of predictability and protection of users. Naturally, it is possible for users to misuse the information (e.g. to abuse the information). Since abuse by users (engine hacks) should be prevented, prior notification may be abstracted to a certain degree.

3.3 Output

3.3.1 Performance Accuracy and Limitations

This section describes the accuracy and limitations of the AI model's result output.

For example, by describing how reliable the results are in terms of significance probability, error rate, and correct answer rate, users can adjust their expectations and use of the AI appropriately.

It would also be useful for users to describe common errors and explain cases that are prone to incorrect answers.

3.4 Overall system design (System)

Depending on the learning method used, AI may produce output with logic that is beyond the comprehension of ordinary users. Whether the AI product can be trusted or not depends on the transparency and accountability before and after using the AI, as well as on the entity developing the AI product and its governance system. Therefore, the section on overall system design and human involvement describes the human involvement as a subject using AI, its design concept, and its governance system, in order to gain understanding of and trust in the AI.

3.4.1 Purpose of Use

Describe the purpose of using AI in the service or system to be provided.

The purpose of use should include the background of development and future vision, which will lead to a more concrete understanding of the business's intention and philosophy of use.

3.4.2 Benefit or Impact

Describe the benefits or impacts of using the AI.

This item can be described from various perspectives, such as from the user's perspective, from the business operator's perspective, or from the perspective of society as a whole. By describing mainly the positive impacts that can be expected, apart from the risks, it is possible to lead to a better understanding of users and society as a whole.

3.4.3 How the AI will be used

Explain how AI is used in the services and systems you provide.

How AI is used in the overall service often occurs in ways that are not obvious to users at first glance. Specifically, it is desirable to explain in detail which processes use AI, and whether the output of AI merely serves to support human decision-making, or whether the AI output itself making the decision.

For example, when AI is used in credit screening based on credit card payment

history information, purchase history, etc., whether the screening result by AI itself becomes the final screening decision or whether a human makes the screening decision while referring to the screening result is completely different in meaning, and unless the results are disclosed, users will not know.

3.4.4 Risk Management

Describe how to assess, minimize, or control risks that may arise from the use of the AI.

Rather than concealing the risks generated by AI, the potential risks should be shared with users, and efforts should be made to minimize them or to control them, in order to coexist with the risks and generate innovation. Such a stance would allow users to use the service with an awareness of the risks, and would also lead to risk management by the users themselves.

It may also be possible to differentiate the company's AI by explaining the advantages of the company's AI in terms of risk management against the risks that similar AIs generally encompass.

Risk management can be organized in terms of data, models, outputs, etc., to make it easier to understand.

3.4.5 Education System

In order to avoid, as much as possible, any damage to users or third parties or new problems to society caused by the use of AI, explain how the business has established an education system for the ethical awareness and points to keep in mind required for the development and provision of AI, the content and frequency of training, and the certification system. The content and frequency of training, certification systems, etc. will be explained. This will help visualize the quality level of handling that the service provider is aiming for.

3.4.6 Audit System

Describe what kind of audit system is in place for data, models, outputs, etc. Identify what kind of audit is conducted after the model is utilized and what kind of feedback is provided in terms of audit entity (internal or external), audit target (audit of the development process or audit of the technology itself), etc.

3.4.7 Control by user

3.4.7.1 Choice of AI Algorithm Use

This section describes how users can choose whether or not to use an AI algorithm, or which AI algorithm to use if there are multiple options.

For example, a social media platform makes a distinction between feeds based on recommendation engines and feeds based solely on posting times, and allows users to choose between the two. However, it is not always clear at a glance whether the distinction is made and whether the user can make a choice or not. If we want to enable users to select AI algorithms proactively, autonomy must be properly communicated.

3.4.7.2 Feedback

Describe whether there is a method of feedback by users to AI algorithms and data, and if so, how it is used.

For example, some aggregation platforms allow users to select items of interest and thereby provide feedback to the AI algorithms. In addition, some providers allow users to correct or add to the profiling results inferred by the provider, so that the AI algorithm can be adjusted to be more optimal.

3.4.7.3 Data Opt-Out

This section describes whether there is a way for users to request to stop the use of their own information, their own intellectual property, etc., in the training data used for AI models, and if so, how to apply for such a right.

Services and systems that assume the existence of a certain number of users who do not want their data to be used for learning data, which is often a problem especially in generative AI, should consider implementing such an

opt-out right. While opt-out rights for the use of personal data are recognized in the EU and some states in the U.S., the use of data beyond personal data must be determined by the business.

3.5 Human Involvement

3.5.1 Persons responsible for development and management

The person(s) responsible for developing and providing the AI and the person(s) responsible for its management shall be listed. This statement clarifies where responsibility for the AI should be attributed. This is in line with the international trend of making the clarification of parties responsible for handling personal information mandatory.

3.5.2 The development team's DEIB policies

Describe the operator's policies and views on "Diversity, Inclusion, Equity, and Belonging" i.e. diversity, fairness, comprehensiveness, representativeness, and inclusiveness taken into account by the team developing the AI.

One way to minimize AI bias is to ensure the diversity of the team that develops and delivers it, or to be consistent with the population of expected users. However, it is rare for a development team to have the necessary and sufficient attributes in all these aspects. Therefore, the policy of the provider could be explained in terms of how it tries to eliminate unconscious bias, and whether a system is in place to check for multiple perspectives.

3.5.3 Management of contractors

When AI is developed and provided together with external contractors, the management methods and systems for such development and provision should be described. For contractors, it is required to describe the measures taken to ensure appropriate development, such as requiring each business to develop AI according to its own guidelines.

3.5.4 Operator run monitoring and maintenance

In particular, an active system for monitoring and maintenance by service providers to provide quality or safety evaluation of AI produced output should be described.

In the actual operation phase, an active monitoring system for output may be required from the viewpoint of trust and safety in order to minimize the impact of erroneous output or illegal and/or harmful information about users. In addition, sharing the specific operational status (e.g., number of detections) with users and the general public, and explaining how to provide feedback about such information to the AI will lead to further trust building.

3.5.5 Inquiry handling

Describe information on inquiry contact points and complaints against output by users and researchers, customer experience system and its operational results (e.g., contents and frequency of inquiries), etc.

For example, in disclosures based on Japan's DPF Transparency Act, Amazon, Google, and others submit in the form of annual reports the types of complaints, the average length of time for processing complaints, and a summary of the results of processing (percentage of results processed, etc.)[20]. This is a useful quality disclosure that shows how each operator is dealing with user inquiries.

3.6 Points to be noted

Regarding points related to human/AI interaction, in particular, some business entities that provide multiple services or products may have a unified system across the board, not by service or product unit, but by business entity unit. In addition, there may be areas

[20] Ministry of Economy, Trade and Industry, "Compilation of 'Assessment of Transparency and Fairness of Specific Digital Platforms'," https://www.meti.go.jp/press/2022/12/20221222005/20221222005.html (viewed January 28, 2023).

in which transparency can be further improved, or disclosure may be required from the perspective of the business as a whole. However, this toolkit, as ver. 1.0, focuses on a single service or system for the time being.

4 Perspectives for Using the Toolkit

As described above, explanations have been provided for each item in this toolkit. However, as repeatedly stated, not all entities are required to disclose all of these items in detail every time. Important perspectives in deciding which items to disclose include the risks inherent in the AI to be used, the entities using the AI, and the stakeholders who need the information.

4.1 Classification by risk/impact of AI use

The EU has classified AI into four risk-based categories in its AI Regulation[21], and does not require uniform regulation and transparency for all AI. Naturally, the transparency required will differ depending on the size of the impact.

Although this is only one reference, for example, the higher the level of transparency required for Tier 1, the more transparency would be expected in the following classification. However, depending on the type of AI used by Tier 1, there may be issues that cannot be resolved by transparency (e.g., when significant damage is caused once the risk becomes apparent), so this point should be kept in mind.

	Tier 1	Tier 2	Tier 3
risk	Those that endanger the life or body of the user or the fundamental human rights and liberal democratic values shared by society.	Any economic impact on users	Other AI (that assists users in decision making or provides operational support)
concrete example	AI to make medical decisions, large scale video surveillance AI, vote destination matching, etc.	Job matching AI for job seekers, contract reviewal AI, image generation AI (for creators), business recommendation engine for restaurants and hotels, etc.	Work shift coordination AI, search engines, etc.

4.2 Classification by entity using AI

The transparency required will increase depending on the entity that uses the AI. In particular, when an AI is used by a public organization in its business, it is the citizens who are affected by its administrative services, and a high level of information transparency is required regarding its use. In addition, a high level of transparency is also required for major platform operators with a dominant position in a certain market in order to ensure the fairness of their transactions.

	Tier 1	Tier 2
core	Government and Public Organizations, Major Platforms	Other

In fact, the Algorithmic Transparency Recording Standard published by the UK

[21] The European Union "Regulatory framework proposal on artificial intelligence" https://digital-strategy.ec.europa.eu /en/policies/regulatory-framework-ai, viewed 28 January 2023).

government at the end of 2022 ([22]) comprehensively organizes how the public sector, including government, should disclose information when using algorithmic tools. Like this paper, it does not concentrate solely on models, but also expresses the accountability that the public sector should fulfill from various perspectives, from data and output to human involvement. It will be referenced in various places in the future as a standard for transparency in light of the importance of the public sector as an entity.

4.3 The information recipient's perspective

The items to be disclosed from the perspective of transparency of the AI algorithm using this toolkit and the manner in which they are described must take into account the different stakeholders who will review the disclosed results. This is called the reader's perspective. This is because each stakeholder has a completely different understanding, need, and use for transparency, and the method of explanation should be individually selected accordingly. Table 2 shows a selection of previous studies that pointed out that each stakeholder has a different understanding of transparency regarding robotics and AI.

stakeholders	Types and Objectives of Transparency
developer	Understand how the system works for debugging and improvement purposes Facilitate monitoring and testing for safety standards
user	Understand what the system is doing and why in order to anticipate what might happen in unforeseen circumstances, and to build confidence in the technology. Understand why a particular prediction or decision was made, ensuring that the system functioned properly, and allow for meaningful challenges (e.g., approval of a review or a criminal conviction).
community	Broadly understand the strengths and limitations of the system, gain familiarity with it, and overcome a reasonable fear of the unknown.
Experts and regulators	Provide the ability to audit in detail the traces of predictions and decisions, especially if something goes wrong (e.g., a collision by an autonomous vehicle). This may require storing key data streams and tracking each logical step, facilitating the assignment of accountability and legal responsibility.
deployer (deployer)	Ensure that users are comfortable with predictions and decisions and continue to use the system. Induce a user to take some action or behave in some way. For example, Amazon recommends a certain product and describes it in such a way as to induce the user to click and purchase it.

Table 2 Differences in each stakeholder's understanding of transparency regarding robotics and AI (adapted from Weller, 2017, translated by the author from[23])

As this table reveals, each stakeholder has different objectives and different levels of understanding, and these differ for developers, users, society in general, or experts and regulators. The selection of items in this toolkit and the manner in which explanations

[22] GOV.UK "Algorithmic Transparency Recording Standard" https://www.gov.uk/government/publications/algorithmic- transparency-template (viewed January 28, 2023).

[23] Weller, A. (2017). Transparency: Motivations and Challenges. arXiv:1708.01870. https://arxiv.org/pdf/1708.01870.pdf (viewed 17 January 2023).

are provided should be tailored accordingly.

It is also important to consider the perspective of the country/region in which these stakeholders are located. Legal regulations, social norms, and value consciousness differ greatly from country to country and region to region. Accordingly, there will be different levels of transparency that should be ensured in accordance with the issues that society is aware of. This toolkit is also intended to be used by internal audit and risk compliance departments as another stakeholder group. By using this toolkit as a checklist, it will help promote in-house understanding of AI, as well as risk assessment and management of the AI in question.

4.4 Granularity and Degree of Transparency

Transparency does not mean that it is sufficient to simply disclose each item in this toolkit. Some researchers have criticized transparency as meaningless or even harmful[24], and one of the strong grounds for this criticism is that there are limits to users' understanding of the disclosure items. Explanations that are too vague or that skirt the details do not contribute at all to gaining trust and improving understanding, which is the goal of transparency. On the other hand, it is also difficult to agree on whether detailed disclosure will achieve these goals, since there is a limit to users' understanding. However, Table 3, which summarizes several previous studies on the granularity and degree of transparency of AI algorithms, suggests a realistic possibility.

Research topics	Transparency result indicators	Findings
Movie recommendation systems	-Acceptance of the system -Performance	Mixed results: no effect on performance, but positive effect on acceptance
Music recommendation systems	-Satisfaction with the system -Reliability	Positive effects on satisfaction and trust
Social robots	-Liability attribution -Attribution of credit	Weak effect
Cultural property recommendation systems	-Acceptance of the system -Reliability -Competence	Weak effect: no effect on trust and competence, partial effect on acceptance
Action recognition systems	-Comprehension -Reliability -Performance	Explaining why or why not improves understanding, confidence, and performance
Music recommendation systems	-Mental model -User confidence	A sound and complete explanation is best for understanding and trust
Facebook's news weed algorithm	-Degree of initial surprise, anger, or frustration -Satisfaction with the level of gradations.	Knowledge of the existence of the algorithm produces positive and negative effects.

[24] Ananny, M., & Crawford, K. (2018). Seeing without knowing: limits of the transparency ideal and its application to algorithmic accountability New Media & Society, 20(3), 973 -989. https://doi.org/10.1177/1461444816676645 (viewed 17 January 2023)

Peer evaluation in MOOCs	-Credibility	The effectiveness of transparency depends on users' expectations and whether they are disobeyed. If the results do not meet the users' expectations, the level of trust will be low, even if transparency is high.
Environmentally friendly mobile apps	-Reliability -Recognized controls	Positive effect on confidence but no effect on perceived control
Online advertisement	-Reliability -Spookiness -Satisfaction	Explanations that are too specific and general lead to creepiness. Intermediate explanations increase trust and satisfaction. Algorithmic transparency leads to disillusionment.
Facebook's News Feed algorithm	-Recognition -Correctness -Interpretability -Accountability	Highest effect on awareness, followed by effect on accountability

Table 3 Prior studies on the impact on individual users on the transparency of various AIs (adapted from Felzmann et al.,2019, translated by the author from[25])

What the previous studies show is that, depending on the type and role of the AI, the granularity and degree of transparency must be devised, which in some cases may rather reduce users' trust, satisfaction, and performance. In particular, it is important to note that if we provide too much information in an attempt to explain in detail, and as a result, the users do not understand the information, it will rather promote a sense of weirdness toward AI, or that seeming transparent, it may lower the level of trust, depending on the behavior of AI. On the other hand, simply putting overly general explanations as an alibi will not produce any positive effects, including PR effects. From the perspective of users and other stakeholders, the degree of disclosure necessary to maximize effects should be considered depending on the type of AI and its role.

5 Use Case

5.1 Purpose of the Use Cases

We would like to note two brief use cases to illustrate how this toolkit can be used specifically. These use cases by no means represent ideal disclosure cases, and intentionally include controversial items. We hope that the use cases will help you consider the selection of what to disclose and how to describe it. Please note that all use

[25] Felzmann, H., Villaronga, E. F., Lutz, C., & Tamò-Larrieux, A. (2019). Transparency you can trust: transparency requirements for artificial intelligence between legal norms and contextual concerns. Big Data & Society, 6(1). https://doi.org/10.1177/2053951719860542 (viewed January 17, 2023).

cases are fictitious cases, and there is no particular system or service in mind.

5.2 When AI is used by a government agency to review applications for grant programs

In this use case, the following points are factors to be considered in achieving transparency of AI algorithms, and the items, method, and degree of disclosure should be addressed in light of them.

(1) Entity.

XX Agency (government agency), and transparency is strongly required.

(2) Risks/impacts of using AI:

Economic impact on the user (grant application or not).

(3) Recipients of information:

The public at large, with wide variations in literacy.

In such a case, a wide range of information is required to be disclosed, and it is necessary to devise a method and degree of disclosing information that is understandable to the public at large, while requiring strong transparency.

An example of disclosure for this use case is provided in Appendix 2. This is by no means an ideal case, and for the sake of convenience of explanation, parts of the disclosure that may require special attention are highlighted. The following is a brief discussion of some of the items that require special attention:

“Benefit/Impact" should be described as concretely as possible from the perspective of each stakeholder in order to gain the understanding of users. Especially in the case of use by government agencies, cost-effectiveness of using taxpayer funds should also be included.

In the " AI integration " section, it should be emphasized that AI does not make decisions, but merely supports humans. It is important that it is explicitly stated that humans are ultimately responsible for review.

The more sophisticated the "DEIB policy of development team" and "DEIB policy of training data" are for services that are intended for use by the public at large, the more sophisticated the policy should be. In particular, consideration should be given to describing in more detail than is currently described about measures to prevent discrimination (different treatment that cannot be reasonably explained) based on a given attribute.

In the " Parameters/features," five major parameters are listed. Then, a brief description is placed for each item. For the parameter "Age and gender," which is particularly prone to misinterpretation as irrational discrimination by the administration, there should be a description that directs users to a separate page for more details. In cases where differences in literacy among users are expected, such hierarchical descriptions are preferred.

As for "Performance accuracy and limitations" and "Monitoring and maintenance by operators," as long as government agencies use the service, they are required to explain in advance the accuracy of AI algorithms and how to respond to possible erroneous judgments, so that users will not have doubts about the results of the assessment.

For AI algorithms that are used by government agencies and whose results have economic impact, it is conceivable that basically all matters should be disclosed in an easy-to-understand manner, except for those that are not allowed to be disclosed due to their nature (e.g., training data including personal data). As mentioned above, the U.K. has taken the lead in initiating transparency initiatives by government agencies, and we will closely follow the practical measures taken in the future.

5.3 The Use of AI Recommendation Functions for Social Media Feeds

Next, we examine the transparency of AI algorithms used for social media feeds by private operators. The disclosure examples in Appendix 3 show that, unlike the use cases in the previous section, there are some items that are not disclosed, while many others are left to the user to control from the viewpoint of usability.

The "source of training data" specifies the repository where sample data is stored. Even if it is not possible to disclose all of the training data that gives operators a

competitive advantage, the release of a certain amount of sample data extracted through random sampling, etc., should be considered as a realistic option.

The "Parameters and features" are described in some abstracted categories for the main parameters, and the relative importance of the parameters is listed in the "Weight". It is also characteristic that the report specifies "as of X, 2023." Another feature is the design that allows users to select an individual post and learn the reason it was displayed, along with its parameters. This type of implementation is gradually starting to appear in major social media, and it is expected to be actively used not only in social media, but also in rating services and matching services.

Various descriptions of "control by users" are included. Ideally, the UI/UX should be designed so that users can notice and control feeds while using them without having to check such items.

6 Regulatory Status in Each Country vs. this Toolkit

The EU has taken the lead in discussing and introducing regulations on AI algorithms, especially in GDPR, P2B Regulation, AI Regulation, DSA, DMA, etc. It can be seen how important it is for the EU to consider human rights and democracy as values for the coexistence of AI and human society. GDPR is unique in that it raises data rights to the discussion of human rights, and AI Regulation adopts a risk-based approach, in which different levels of regulations are applied according to the type of risks involved with AI. This paper takes the same position in classifying the level of transparency according to the entity that develops and provides AI, who uses the AI, how it is used, and the risks it poses. P2B Regulation, DSA, and DMA, in particular, focus on the significant role that platform operators and online service providers play in society, and apply different levels of transparency according to size P2B Regulations. DSA and DMA, in particular, include regulations that require the implementation of appropriate communication with society from the perspective of transparency, depending on the size and type of business.

Meanwhile, in Japan, the Act on Improving Transparency and Fairness of Digital Platforms(TFDPA) has also introduced a joint regulation requiring transparency of AI algorithms for certain designated major platform operators. As already mentioned in this paper, voluntary disclosure by platform operators has been realized, and platform users' understanding is also increasing.

Note that the U.S. still does not have specific regulations in place regarding AI algorithm transparency, but discussions have already begun in Congress. The most direct example is the Algorithmic Accountability Act ([26]), which also calls for a high level of transparency and accountability for high-risk cases such as AI responsible for important decision-making. However, this bill has not yet been passed, and the U.S. Congress will continue to debate this point of view for the time being.

While we will not go into the details of each regulation in this report, we have prepared a table in Appendix 4 that compares the regulatory status in each country with this toolkit. Of course, each law and regulation has a different regulatory target and different regulatory content, so the comparison table is not intended to explain the interpretation of individual laws and bills, but only to show, in broad terms, what items are of interest in each law or bill. We hope you can see from the comparison table that this toolkit is calling for a bird's eye view and systematic discussion on the topic of transparency.

7 Conclusion

In order to maximize the positive impact of AI innovation, it is essential to design and operate technical, organizational, and social systems that enable stakeholders to recognize the risks of AI and to adjust their interests appropriately and flexibly. In other words, an agile governance framework for AI. This paper challenges the overarching and systematic

[26] CONGRESS.GOV "H.R. 6580 - Algorithmic Accountability Act of 2022" https://www.congress.gov/bill/117th-congress/ house-bill/6580 (viewed January 28, 2023).

organization of the most controversial governance approach, transparency of AI algorithms, and proposes a practically applicable toolkit. We should avoid hampering innovative AI in social implementation with misunderstandings caused by lack of communication, and the role that transparency can play in this regard is not small. However, if regulations and norms are formed each time a problem arises in the absence of a systematic organization, regulations will be repeatedly added without a blueprint, which lacks predictability and "transparency" will become a self-objective only for formality apart from the original purpose. As a result, innovation will be stifled. This is why such regulations must be systematic based on a unified concept, while pursuing generality, clarity, and flexibility with room for discretion so that business entities and government agencies can actually apply them.

In this paper, we have constructed a toolkit as a collection of systematic disclosures and examples, taking into account not only regulations proposed by national authorities but also a wide range of new risk events arising from various AIs, as well as prior research on AI algorithms. We also incorporated the viewpoint that the transparency degree that a business entity or government agency can devise can have a positive impact not only on its social credibility, but also on different indicators such as satisfaction. In addition, the toolkit was intended to be highly convenient for business entities and government agencies responding to self- and co-regulations by putting the disclosures in a list format and making it discretionary and selective in terms of AI algorithm providers, users, risks, and other factors. We would be more than happy if business entities and government agencies could use this toolkit for communication with users, society, authorities, and experts, or for internal risk management. We believe that this is one of the best agile governance practices to maximize the impact of AI innovation.

This toolkit is version 1.0 and we will continue to update it based on the points raised in future discussions. We would be grateful for guidance from readers in pointing out excesses and deficiencies.

8 Acknowledgments

Finally, we would like to express our sincere gratitude to Dr. Naoko Munakata, Dr. Hiroki Habuka, the University of Tokyo School of Public Policy, and the students of the Innovation Governance Expert Program for providing us with the opportunity to propose this paper and for their continued useful suggestions from various perspectives.

Appendix 1 List of Transparency Disclosure Items for AI Algorithms

Lv. 1	Lv. 2	Main entry (in dictionary)	System	Data	Model	Output	Human	Disclosure Items
Product concept	Purpose of use	3.4.1	○.					Purpose of using AI
	Benefit/Impact	3.4.2	○.					Assessment of the benefits and impact of using AI
	AI integration	3.4.3	○.					Processes where AI is used, and the role AI plays (e.g., does it make decisions itself or does it support human decision-making?)
Development team	Development manager and operation manager	3.5.1					○.	Department/person responsible for development and operation
	DEIB Policy of development team	3.5.2					○.	Diversity of the development team and alignment with the expected users
	Outsourcing management	3.5.3					○.	Management system and methods of outsourcing
Data	Overview of training data	3.1.1		○.				Overview of training and test data
	Sources of training data	3.1.2		○.				Location of training and test data
	Training data collection	3.1.3		○.				How and where to obtain training and test data
	DEIB policy of training data	3.1.4		○.				Diversity of training and test data, consistency with expected users
	Actual operational data	3.1.5		○.				Overview of the actual data used (what does AI actually use as Data to extract Outputs)
AI algorithm	Utilization Phase	3.2.1	○.					Phases of algorithms used (research, testing, production, etc.)
	Learning Methods	3.2.2			○.			Learning methods of AI (expert systems, rule-based, machine learning, deep learning, etc.)
	Explainability	3.2.3			○.			Explainability of the algorithm
	Parameters/features and their purpose/reasons	3.2.4			○.			Parameters abstracted and verbalized to a reasonably understandable degree

Appendix 1 List of Transparency Disclosure Items for AI Algorithms

	Weight and its purpose/reason	3.2.5			○.			Weight for main parameters and why they are relatively important compared to other parameters
	Different treatment for each user category, including billing	3.2.6			○.			Relationship between parameters, weighting and user classification such as actions and attributes including billing
	Due process for change	3.2.7	○.					Due process for changing the algorithm, impact of the change, purpose of the change, etc.
Output	Performance accuracy/limitations	3.3.1				○.		Significant probability of output, error rates, false negatives/false positives, etc.
	Monitoring and maintenance by operators	3.4.4					○.	Monitoring and maintenance system for output quality (frequency, method, and system), number of detections
Control by user	Choice of algorithm use	3.4.4.1	○.					The right to choose whether or not to accept the algorithm
	Feedback	3.4.4.2	○.					Availability, acceptability, and method of feedback on algorithms and datasets by users
	Data opt-out	3.4.4.3	○.					Availability of opt-out method from training data
Management structure	Risk management	3.5.5	○.					Assessment, minimization, or management response to the risks of using AI
	Educational system	3.5.6	○.					Status of education related to the development and provision of AI
	Audit system	3.5.7	○.					Audit system and implementation status for the development and provision of AI
	Inquiry response	3.5.8					○.	Contact person/appeal office, customer experience system (frequency, methods, and structure), number of complaints after a certain period of time

Appendix 2: Use Case 1: Government agencies use AI to review applications for grant programs

Lv1	Lv2	Disclosure Example
Product concept	Purpose of use	The department in charge of reviewing grant applications will use an AI algorithm (the "AI") to scrutinize a large number of review documents and identify those applications that do not meet the screening criteria.
	Benefit/Impact	Benefit to the department in charge of examination: The department plans to conduct 2.5 million examinations per month, and the use of this AI will reduce the man-hours required for examination by 70%. Therefore, the number of examination days will be shortened by approximately one week and the monthly labor cost will be reduced by approximately 420 million JPY compared to the case where all examinations are conducted by humans. Benefit to the applicants: The use of the AI will shorten the time to receive the grant by approximately one week.
	AI integration	The AI will be used in the following examinations. However, in all examinations, the final decision is made by the examiner, and the role of the AI is only supplementary sorting for that decision. Formal examination of documents: Formal aspects such as erroneous or unclear statements are examined. Formal screening of eligibility criteria: The applicant will be screened for eligibility for the grant based on the business income, type of industry, etc. as filled in the documents. Substantive review of eligibility criteria: Priority of grant recipients to be distributed within a budget cap is reviewed from a variety of data (see below).
Development team	Development manager and operation manager	CTO of XX Agency (concurrently serving as head of Information Systems Division): XX
	DEIB Policy of development team	Recognizing the diversity of applicant demographics (particularly age, gender and industry), XX Company, the developer of the AI, has established a process during the testing phase whereby subjects of each demographic (those conducting the pilot application) are checked for the occurrence of discriminatory results.
	Outsourcing management	We require XX, the developer of the AI, to comply with the development guidelines set forth by Japan and, in particular, require reports on the following items: If the results are unexplainable or extremely difficult to explain If there is any doubt about the accuracy of the AI's judgment XX.
Data	Overview of training data	The data used to develop the AI in this case were past grant application documents held by the Agency, past review results and reasons, government statistical data including earnings by year for each industry, data related to tax returns, and data related to tax payments.

	Sources of training data	(This is sensitive personal information and the data cannot be disclosed.)
	Training data collection	Data submitted directly by the individual to the Agency or to another agency that holds the data.
	DEIB Policy of training data	As mentioned above, we use data from past grant applications of the same purpose as training data, but we exclude from the data set any outlier data (e.g., industries that were explicitly excluded) that we consider not to constitute the population of the current study.
	Actual operational data	The data submitted by the applicant for this grant application will be used in the review process.
AI algorithm	Utilization phase	The AI is already in the production phase.
	Learning methods	Machine learning is the main method, but there are some rule-based aspects of formal screening.
	Explainability	Basically, the results output by the AI can be explained by the following parameters, features, and weight. However, there are extremely rare cases that are difficult to explain, in which case the AI output results will not be used and the examiner will make a decision on a case-by-case basis.
	Parameters/features and their purpose/reasons	In particular, the substantive examination of the subject criteria has the following as key parameters/features: • Positive results are more likely to be generated for industries that are considered to have experienced a deterioration in earnings this fiscal year based on multiple government statistics, etc. • Age and gender (in light of past screening results, positive results may occur as attributes that may be particularly protective due to their relationship to the industry and earnings situation. For more information, please click here.) • The current year's earnings (positive results are more likely to be generated if the average of the current year's earnings is lower than the average of the previous five fiscal years). • Amended tax return history for the past three fiscal years • Tax payment status for the past 10 fiscal years (negative results are likely to be generated only if there are delinquent taxes or additional taxes due).
	Weight and its purpose/reason	Industry and the current year's earnings situation are the most important factors.
	Different treatment for each user category, including billing	N/A
	Due process for change	There are no plans to change the algorithm of this AI within the fiscal year.
Output	Performance accuracy/limitations	When the AI results were compared to the past review documents and review results, it was 97% accurate. As for the remaining 3%, it can be said that many of the applications contain outliers or there are doubts about the validity

		of the past examination results. Therefore, multiple examiners will review the results of the AI, and the final decision will be made by humans.
	Monitoring and maintenance by operators	The results of the AI will be reviewed alternately by two reviewers, and if both reviewers find that they disagree with the AI, the results will be feedback to XX Company, the developer of the AI, for use in improving the AI.
Control by user	Choice of algorithm use	None, due to the nature of the usage of this AI.
	Feedback	None, due to the nature of the usage of this AI.
	Data opt-out	None, due to the nature of the usage of this AI.
Management structure	Risk management	There is a risk that the AI in this case will output inappropriate screening results, and this is addressed below: - Review and feedback by multiple reviewers - Check for consistency with past review results - Feedback on appeals by the applicant and any errors in the AI that are revealed by such appeals.
	Educational system	The reviewers have received a total of 15 hours of training in the specifications and use of the AI and have passed the standard test required by the Agency.
	Audit system	The Agency's audit department will conduct an audit of this year's grant application decisions within one year, and any doubts about the AI or the reviewing officer's decisions may be made public and will be feedback to Company XX, the developer of this AI.
	Inquiry response	Inquiries or appeals regarding this AI or the grant application should be directed to XX Agency XX Grant Application Inquiry Office Phone number: XX E-mail address: XX

Appendix 3: Use Case 2: Using AI Recommendation Functions for Social Media Feeds

Lv1	Lv2	Disclosure Example
Product concept	Purpose of use	Recommendation AI (the "AI") will be used within the social media "YY" post feed to provide users with content that is highly convenient and relevant to them, rather than displaying content in the order of breaking news.
	Benefit/Impact	Users receive more satisfying content and spend their limited time optimally.
	AI integration	-
Development team	Development manager and operation manager	-
	DEIB Policy of development team	The development team is guaranteed to be diverse in terms of race, gender, religion, and education, and at least one person is required to review the development of this AI from a minority perspective.
	Outsourcing management	-
Data	Overview of training data	The training data includes the past 3 years of posts in YY, user reactions to the posts (reaction buttons, comments, diffusion, viewing time, etc.), and post formats (text, image, video).
	Sources of training data	Sample datasets are available from the following repositories: YY.com/repository/feed/sample-dataset/
	Training data collection	Feed in YY
	DEIB Policy of training data	Since the data is obtained from YY internal feeds, a certain degree of representativeness of YY users is ensured. For example, users who stopped using the system after the first startup or users who behave like bots are excluded from the population of the dataset as outliers.
	Actual operational data	Actual post data by the user is used for actual operation.
AI algorithm	Utilization phase	-
	Learning methods	-
	Explainability	-
	Parameters/features and their purpose/reasons	The AI in this case mainly holds the following parameters. In addition, if you select "Why is this post displayed?" in the detail menu of each individual post, the reason for displaying the individual post will be explained along with the parameters used. · Relationships among users · Amount of common friends among users · Content of user interactions (e.g., mutual reactions)

Appendix 3: Use Case 2: Using AI Recommendation Functions for Social Media Feeds

		· Amount of messages between users · Time spent by one user on the other user's post · User classification · Post Content · Format (text, images, video) · Being current · Relevance · Regional character · Reactions to the post · Amount of reactions · Frequency of reactions · Degree of reaction cross-community
	Weight and its purpose/reason	The parameters of relative importance as of X month 2023 are the amount of relationships and reactions among users and the degree of cross-community.
	Different treatment for each user category, including billing	Payed users' posts are preferentially displayed more often and at a higher frequency than non-payed users' posts.
	Due process for change	For material changes to the AI, we will provide at least 7 days' notice to the user. For adverse changes related to payed users, we will provide a reasonable period of time to cancel the recurring payments.
Output	Performance accuracy/limitations	-
	Monitoring and maintenance by operators	To ensure that the AI does not handle too much illegal and harmful information, our Trust & Safety team regularly monitors the feeds and removes or down-ranks the posts in question. The nature and number of these actions are published in our annual report each fiscal year.
Control by user	Choice of algorithm use	If you would like to enjoy YY feeds in the order of breaking news, not in the order of recommendation by the AI, please select "Settings" →"Change Display Order" →"View in Order of Breaking News." Likewise, you can always enjoy the feeds in the order recommended by the AI from the same setting.
	Feedback	If you find a post in YY's feed that you are not satisfied with, please give feedback to the AI by long-pressing the post or clicking the "..." button below the post and then clicking the "I am not interested in such post" button. You can also select a reason on the next screen to make a more personalized feedback. Please note, however, that there is a time lag before the AI can make good use of your feedback.
	Data opt-out	To exercise your rights regarding your personal data, please go to "Settings" →"About Personal Data."

Management structure	Risk management	-
	Educational system	-
	Audit system	-
	Inquiry response	For inquiries about this AI, please contact support-recommendation@example.com

Appendix 4: Comparative Table of Regulatory Status in Each Country and this Toolkit

Lv1	Lv2	GDPR (EU)	AI Regulation (EU)	P2B Regulation (EU)	DSA (EU)	DMA (EU)	ATRS (UK)	AAA (US)	TFDPA (JP)
Product concept	Purpose of use	○	○	○	○		○		
	Benefit/Impact		○				○		
	AI integration	○	○				○		
Development team	Development manager and operation manager		○		○		○		
	DEIB Policy of development team								
	Outsourcing management						○		
Data	Overview of training data	○			○		○		
	Sources of training data				○	○	○		
	Training data collection	○					○		
	DEIB Policy of training data								
	Actual operational data	○				○	○		
AI algorithm	Utilization phase						○		
	Learning methods						○		
	Explainability			○	○				○
	Parameters/features and their purpose/reasons			○	○				○
	Weight and its purpose/reason			○	○				○

Appendix 4: Comparative Table of Regulatory Status in Each Country and this Toolkit

	Different treatment for each user category, including billing			○					
	Due process for change			○			○	○	○
Output	Performance accuracy/limitations						○		
	Monitoring and maintenance by operators		○		○		○		○
Control by user	Choice of algorithm use	○		○	○	○	○	○	
	Feedback			○			○		
	Data opt-out	○				○			
Management structure	Risk management		○		○		○	○	○
	Educational system		○		○			○	○
	Audit system		○		○			○	○
	Inquiry response	○			○		○		○

*GDPR: General Data Protection Regulation *AI Regulation: AI Regulation *P2B Regulation: The EU Regulation on platform-to-business relations
*DSA: Digital Services Act *DMA: Digital Market Act *ATRS: Algorithmic Transparency Recording Standard
*AAA: Algorithmic Accountability Act *DPF Transparency Act: Act on Enhancing Transparency and Fairness of Certain Digital Platforms