Towards an AI Governance Framework (ver1.0)

July 2024

Digital Policy Forum Japan

Basic Concept

Technological developments surrounding generative AI are evolving at a remarkable rate, and the implementation of generative AI into socioeconomic systems is accelerating.

Under these circumstances, the creation of rules to control AI is rapidly steering away from ideological discussions to concrete and specific ones. AI laws have been passed in Europe[1] and China[2] , and in the United States, specific governance rules are being considered based on Presidential Executive Order[3] . In Japan, earnest discussions[4] including a legal framework are about to begin among stakeholders.

In this document, keeping in mind these trends in generative AI, the basic perspectives for consideration are,

- Minimization of the risk of AI,
- Development of an environment that maximizes the utilities of AI,
- Creation of generative AI markets that makes these environments as autonomous as possible

This document will outline issues for the development of an "AI governance framework" or a mechanism to secure the "controllability of AI technology"[5] in order to achieve the three objectives in a well-balanced manner.

In discussing AI governance, it is necessary to constantly compare and contrast the balance between the benefits and risks that AI brings. Among the benefits of AI, the personalization of AI (decentralization of intelligence) may enable individuals to enjoy highly convenient services while technically guaranteeing their sovereignty over their data use. On the other hand, AI risks include the risk of AI manipulating humans and the risk of AI replacing humans. Such risks should be solved technologically as much as possible, and the introduction of more regulations than necessary is not appropriate from the viewpoint of encouraging innovation.

The discussion in this document will be conducted mainly with the current available generative AI for the general public in mind, and will not cover Artificial General Intelligence (AGI) except for some cases. However, as AI technologies are expected to continue evolving exponentially, the content of this document will be updated on an ongoing basis (see "Future Work Plan" at the end of this document).

I. Risk minimization

(1) risk management

There are methods of AI management that divide risks (including negative impacts on human life and basic human rights) into several levels (e.g., AI law in the EU classifies risks into four levels[6]). This is an attempt to manage risks inherent in the AI model, but it requires extensive discussions on how to define the scope of risks to be controlled and on what criteria to rank the risks.

1. It is necessary to consider whether risk assessment should be conducted for AI itself or for each AI provision case (use case) individually.
2. The discipline on AI will be enforced on three groups of entities: "AI developers," "service providers" who incorporate AI into their services, and "end users," rather than a dichotomy between developers and users.
3. It is necessary to discuss how it is appropriate to combine self-assessment and third-party assessment (e.g., audit systems) for risk assessment.

From the perspective of managing risks to AI models, AI vulnerability research or “red teaming” should also be considered for technical criteria and auditing mechanisms, similar to the risk assessment described above.

There is a risk of AI not performing its expected functions or malfunctioning due to data poisoning attacks[7] and others in the AI learning process (cyber attacks against AI). In addition, there is an emerging risk of AI being used to detect vulnerabilities and create malware, generate fake accounts, disseminate false information, etc. (cyber attacks by AI). Specific measures to deal with such "cyber attacks against AI" and "cyber attacks by AI" should be taken promptly.

(2) Regulatory approach and effectiveness

There are several regulatory approaches to AI, including hard law (legal regulations), soft law (self-regulation by the private sector), and co-regulation[8], which falls somewhere in between.

For example, in China and the EU, policy development is oriented toward hard law, while in the U.S., policy development is centered on co-regulation. However, even in the case of hard law-oriented policies, there is a certain range of approaches to discipline, ranging from the "basic law" approach clarifying the philosophy and specifying the roles of each entity in order to realize the philosophy, to "business law" approach to impose specific regulations on the conduct of entities.

In order to address issues such as freedom of expression while avoiding to hamper innovation by laws and regulations, one possible approach is to enact an AI law (hard law), with the national government taking the lead in developing safeguards for AI risks such as technical standards, and adopting a co-regulatory approach [9] for auditing standards and disinformation countermeasures. One possible option is to adopt co-regulatory approach for auditing standards and countermeasures against disinformation.

If the above approach is to be adopted, the following issues need to be considered: (a) whether the regulation should be limited to AI developers as described in (1) above (service providers should not be subject to the regulation), (b) whether the regulation should cover only cases where AI development is conducted as a business (including cases where AI-related business costs are covered by providing other services in an integrated manner), (c) what methods should be used to ensure the effectiveness of the regulation, such as a registration system, and (d) whether players above a certain size should be subject to the regulation in light of their social impact.

(3) Possibility of "model collapse"

In the process of learning data over several generations, AI often goes through a process of discarding data with few occurrences (for the purpose of improving the hit rate for queries). In this case, it has been pointed out that a so-called "model collapse" may occur, in which minority opinions are truncated, resulting in a model that differs from the original AI model may occur.[10] Leaving such a situation unchecked will lead to the proliferation of data

inundated with inaccurate and substandard data and the ongoing contamination of the data space.

To protect the data space from being contaminated by biased intelligence, it is necessary to consider certain rules (e.g., certification system), such as limiting AI training data to those created by humans or clearly indicating to the outside world that the AI is a trained AI.

It would be effective to promote open data policy, where documents whose copyrights have expired or documents created by public organizations are widely available for use as training data.

(4) Product handling

AI takes in training data, forms a model, and outputs data as a product by utilizing the model. Ensuring the integrity of "data not tampered with," (3) above means ensuring the integrity of input values (training data), but efforts to ensure the integrity of output values (products) are also necessary (See Section (5) for AI models that fall between the two).

In the world where a vast amount of disinformation is already circulating using generative AI, it is necessary to consider specific measures on how to combat disinformation effectively while assuming a co-regulatory approach.

An introduction of digital watermarks to certify that AI products conform to copyright law and were legitimately created is considered effective. However, it is necessary to examine international technical standards, standards for entities that issue digital watermarks, and “distributed” digital watermarking from both operational and technical aspects of the system. (a flexible mechanism for mutual recognition of multiple digital watermarking systems), etc., should be examined from both operational and technical aspects of the system.

It should be noted, however, effectiveness of this flexible mechanism requires continual verifications and updating in light of rapid changes in the technological environment.

(Notes)

With regard to items (1) through (4) above, amid rapid technological innovation, some AI-related discussions in the past tended to be far removed from the actual market situation, and to be too abstract or unnecessarily

confirmative.

In discussing the effectiveness of regulations, promotion measures, and user protection, voluntary disclosure of information by AI developers is all the more necessary. It is vital we make changing environments succinct and exposed at all times for fruitful discussions.

II. Improvement of convenience

(5) Prohibition of digital discrimination

As in the case of model collapse ((3) above), AI models may undermine impartiality and neutrality, resulting in unreasonably discriminatory treatment of certain users, or excessive profiling may result in "unintended" disclosure of information that exceeds an individual's sovereignty over the use of their data.

In order to prevent such digital discrimination, we need to arrive at an audit system (self-audit or third-party audit) to ensure fairness and neutrality of AI models. In addition, when providing AI-embedded services to users, the boundary of responsibilities between AI developers and service providers that incorporate AI must be clarified for ensuring user protection.

(6) Active use of AI

A range of initiatives have already been taken to utilize AI. However, given the fact that data utilization efforts are lagging behind in the education and medical fields, it is necessary to aggressively promote AI utilization in these fields in today's Japanese society with drastically declining birthrate and aging of general population.

In particular, a system that links and analyzes relevant data collected, by consent, from students and patients, is expected to improve individualized education programs and medical services.

It is also necessary to consider certain safeguard measures to ensure that such data linkage does not lead to excessive profiling. AI analysis will enable automatic linkage of medical record, for example, where data linkage has not progressed due to existing incongruence in data formats among regions of the country and across organizations.

In addition to the fields of education and medical care, AI should be actively utilized in a wide range of fields, including environmental measures, which are global issues, disaster prevention to protect human life and property, and culture to realize "preferred" lifestyles.

We need to broaden perspectives and deepen understanding on what points need to be taken into account from the viewpoint of active use of AI in these fields and on what technological developments are necessary, what protection measures against privacy are truly effective in avoiding exposures of personal data in imported data.

There is an urgent need for "AI literacy education" for correctly understanding the risks of AI as well as the benefits. To sensitize the youth population, public-private partnership in educational activities to raise awareness of the risks of AI, similar to the efforts for the Internet use should prove useful.

III. Fostering a healthy market

(7) Building a healthy ecosystem

The evolution of AI should basically be driven by the ingenuity of the private sector. The government should actively provide support for this process and furnish the market with policy support for ensuring the public interest.

In doing so, competition policy to establish a healthy market environment is critical to ensure an ecosystem of diverse actors, including developers and users of AI.[11]

It is necessary to establish a mechanism to watch out for possible barriers for entry into AI-related markets and for anti-competitive behavior such as abuse of dominant positions by large companies.

The current leading AIs are mainly provided by large platform providers with massive resources; we must be mindful of possible exploitative abuses in the AI market or its adjacent markets (e.g., platform businesses) by such players in the future.

There is a concern that vertically integrated business models deployed by AI developers or platformers across multiple layers may allow big players to have

market dominance over other AI developers and are more likely to abuse market dominance over adjacent markets as well. Discussions on competition policies are imperative.

It is expected that hybrid networked AI, in which traditional cloud-based and distributed AIs coexist, will become more common, and it is necessary to discuss how to think about the "concentration and dispersion" of such AI.

The EU AI law includes provisions for extraterritorial application of laws, however, consideration must also be given to the possibility that increased extraterritorial application may lead to excessive regulation, where regulations by multiple countries may be superimposed on one country.

(8) Ensuring openness

One of the main reasons for the explosive spread of the Internet is its openness. Similarly, there are two possible business models for to AI: closed proprietary AI and open AI.

As chosen by both Europe and the U.S.[12] ,"openness" is critical for guaranteeing overall qualities of services and a sufficiently competitive environment.

From this perspective, further discussion is needed on issues to be considered from a policy perspective, such as the use of open source, how to ensure interoperability among different AIs, promotion of standardization to realize such an environment, and R&D support based on the premise of encouraging open-type AI development. Further discussion is needed on these issues from a policy perspective.

In advocating increased use of open source data, we need to consider policy-based support for a wide range of R&D including means for interoperability among different AIs through standardization.

Japan already lagging behind in the global market in AI-related technology development and businesses, the government should consider taking proactive measures to promote open-type AI, including support for AI-related ventures.

(9) Fostering international consensus

AI development and services in cyber space are not bound by country

borders.

This means the issues raised above must be dealt with across existing national boundaries. We need to look for opportunities to gain broad consensus on the issues among countries while in each country the steps must be taken to incorporate necessary measures into its existing legal system and rules.

In doing so, it is critical to proceed in a manner that secures sufficient participation from the Global South, in light of the fact that AI has great potential for solving issues faced by the Global South and possibly help these countries “leap frog”.

A particularly urgent task in fostering such an international consensus is the formation of norms for the military use of AI, as proposed at the Conference on Responsible AI in the Military Domain (REALM Summit) held in The Hague in February 2023. And, it is necessary to expand voluntary commitments on the use of AI, as proposed in the "Political Declaration on Responsible Use of AI and Autonomy”. [13] At the same time, incorporating a mechanism for AI security audits (inspections) within the UN security framework is also worth considering. Discussions on the nature of such AI and security should be hastened now the military use of AI has already become a reality.[14]

(10) Addressing Ethical Issues

With the rapid progress of AI, the possibility of "self-conscious" AI in the future needs to be taken into account. Therefore, as in the life science field, ethical issues related to AI research should be properly addressed and specific research ethics regulations and research approval processes should be established.

We need to develop ethical guidelines for issues such as "making AI with self-awareness" over such things as self-reproduction or self-repairing and building in self-discipline for AI based on shared ethical values.

Future Work Plan

The purpose of this document is to provide direction for the discussion of AI-related legislation in Japan and the growing global discussion of AI

governance.

Based on this document, the DPFJ will continue to hold hearings with experts from the three perspectives of AI technology, policy and utilization, and will update this document on an ongoing basis. It plans to take the opportunity of updating this document to hold open forums to broaden the scope and to deepen understanding AI governance.

The DPFJ will seek opportunities for collaborating with other forums that are engaged in similar discussions to foster greater consensus.

The DPFJ hopes to arrive at a framework for AI governance and finalize the document by the end of this year.

---

[1] In June 2023, the European Parliament adopted the AI Act, which is being phased in starting in May 2024, with full application scheduled for summer 2026.
https://artificialintelligenceact.eu/

[2] In August 2023, China enacted the Regulations for the Management of Generated Artificial Intelligence Services. This Regulation (Article 4) prohibits the creation of content prohibited by law or administrative regulations and only allows generated content that "adheres to the core values of socialism".
(Source: Masashi Harada, "China's 'Provisional Measures for the Management of Generated Artificial Intelligence Services and Commentary," Corporate Legal Affairs Navigator (July 21, 2023).
https://www.corporate-legal.jp/matomes/5362

[3] In October 2023, the U.S. government released the Presidential Executive Order on AI governance. The order includes measures to be taken by government agencies, including the establishment of standards for vulnerability research (red teaming), clear guidance on the prohibition of algorithmic discrimination, and support for the appropriate use of AI in healthcare, education, and other fields.
https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe- secure-and-trustworthy-artificial-intelligence/

[4] See, for example, AI Strategy Council, "'Approach to AI Institutions'" (May 2024).
https://www8.cao.go.jp/cstp/ai/ai_senryaku/9kai/shiryo2-1.pdf

[5] Controllability of AI technology" has two aspects: development and use (social implementation). In this document, as the appropriate balance between benefits and risks is to be considered, we take the position that the controllability in actual use (active use with a certain degree of control) is more important. However, in the latter section (10), the discussion includes not only the use aspect but also the development aspect.

[6] The European AI Act classifies AI risks into four categories: unacceptable risk (development prohibited as posing a direct threat to human life or fundamental human rights), high risk (obligation to conduct prior conformity assessment, register in database, etc.), limited risk (obligation to ensure transparency to inform users that they are interacting with an AI), and minimal risk (no regulation).

[7] In a data poisoning attack, the attacker attempts to modify the model to function maliciously by inserting tainted data into the training data that produces incorrect output. In a data evasion attack, noise or other factors that are imperceptible to humans are mixed into the training data to mislead the AI's judgment results.

[8] In the case of co-regulation, the national government presents the basic policy of the rules, businesses that agree with the policy apply the rules based on the basic policy and report the results to the national government,

which evaluates the results and revises the basic policy as necessary. In Europe, this system has been adopted as a countermeasure against disinformation of platformers.

As an example, co- regulation in the AI field, in July 2023, prior to the publication of the Presidential Executive Order (see footnote 3), a non-binding agreement was reached between the Office of the President and seven AI-related companies (Amazon, Anthropic, Google, Inflection, Meta, Microsoft & OpenAI). In September of the same year, a non-binding agreement was reached between the Office of the President and seven AI-related companies (Amazon, Anthropic, Google, Inflection, Meta, Microsoft & OpenAI) to ensure safety, security and trust in AI development. In September of the same year, eight companies (Adobe, Cohere, IBM, Nvidia, Palantir, Salesforce, Scale AI, and Stability) joined this agreement in addition to the above seven companies.

[9] If a co-regulatory approach is adopted, there are concerns about disparities in voice among players and reduced transparency and openness due to the absence of legal regulation, so appropriate public support must be clearly incorporated into the operational policy.

[10] I. Shumailov et al. "The Curse of Recursion: Training on Generated Data Makes Models Forget" arXiv (May 2023)
https://arxiv.org/abs/2305.17493

[11] OECD "Artificial Intelligence, Data and Competition," OECD Artificial Intelligence Papers No. 18 (May 2024).
https://www.oecd.org/daf/competition/artificial-intelligence-data-and-competition.htm

[12] In Europe, the invitation document "Competition in Virtual Worlds and Generative AI: Calls for Contribution," published in January 2024, presents a list of issues related to generative AI and competition policy. The list of issues related to generative AI and competition policy is presented.
https://ec.europa.eu/commission/presscorner/detail/en/ip_24_85

In addition, the U.S. Presidential Decree (see footnote 3) lists "promotion of a fair, open, and competitive ecosystem" as one of the main promotion items from the perspective of encouraging innovation and competition.

[13] This proposal (US DoS "Political Declaration on Responsible Use of Artificial Intelligence and Autonomy" (February 2023)) proposes that military AI The proposal (US DoS "Political Declaration on Responsible Use of Artificial Intelligence and Autonomy" (February 2023)) assumes that military AI will only be used in a manner consistent with its obligations under international law (particularly international humanitarian law), and includes the following elements: publication of principles for the design, development, deployment and use of military AI; implementation of measures to minimize unintended bias; development of auditable military AI; and rigorous testing and assurance of the safety, security and effectiveness of military AI throughout its lifecycle. The agreement includes voluntary commitments by countries to conduct rigorous testing and assurance of the safety, security, and effectiveness of military AI throughout its lifecycle, etc. Currently, 51 countries, including Japan, have endorsed the agreement.
https://www.state.gov/political-declaration-on-responsible-military-use-of-artificial-intelligence-and-autonomy/

[14] According to an investigative report by the Israeli online media "+972 Magazine" in April 2024, the Israeli military is using a generative AI "Lavender" to extract 37,000 people in the Gaza Strip to make a list of operatives, which is then used to target and attack, among other actions.
(Source:) Yual Abraham "Lavender': The Ai machine directing Israel's bombing spree in Gaza" +972 Magazine (April 3, 2024)
https://www.972mag.com/lavender-ai-israeli-army-gaza/

See also Yasunori Kawakami, "Targeting 37,000 people in Gaza: AI machine 'Lavender' revealed," Yahoo! News (April 9, 2024) for more details on the above investigation.
https://news.yahoo.co.jp/expert/articles/c72d4cbc32aa5577eac494dfd75b43652a20555f