# The Self-Checklist
## on National Guidelines for AI Ethics (draft)

2021.12.02

Center for AI & Social Policy, KISDI

**Ahram Moon**
Associate Fellow Researcher

# 1. Background

In December 2020, **the National Guidelines for Artificial Intelligence Ethics** were established as a new social norm to maximize the benefits of the AI information society, to minimize its social risks, and to secure social trust in AI.

» **Present three basic principles and ten key requirements that should be observed by all members of society throughout the entire process of developing and utilizing AI**

  → **Aim to promote social acceptance and trust in new technologies**

» **(Basic Principles)** Respect for Human Dignity, Common Good of Society, Fitness for Purpose

» **(Key Requirements)** ① Human Rights, ② Protection of Privacy, ③ Respect for Diversity, ④ Prevention of Harm, ⑤ Public Good, ⑥ Solidarity, ⑦ Data Management, ⑧ Accountability, ⑨ Safety, ⑩ Transparency

### "AI for humanity"

· Humanity is the highest value that must be upheld by AI

· Developing and utilizing "AI for humanity" requires that:

✓ AI is useful for humans and preserves humanit

✓ AI does not harm the human mind and body

✓ AI promotes personal well-being and positive social changes



⑩ Transparency  ① Human Rights
⑨ Safety  ② Protection of Privacy
⑧ Accountability  ③ Respect for Diversity
⑦ Data Management  ④ Prevention of Harm
⑥ Solidarity  ⑤ Public Good

Respect for Human Dignity
Humanity
Common Good of Society · Fitness for Purpose

# 1. Background

In May 2021, the MSIT announced **The Strategy to Realize Trustworthy Artificial Intelligence**

» **Create a support system to implement action plans for the National Guidelines for AI Ethics and to ensure that the private sector can regain trust in AI through their own efforts**

» The Strategy emphasizes the overcoming of the technological limitations of AI as well as the need for institutional overhaul and the **spread of AI ethics** to prevent the potential risks posed by the misuse/abuse of AI

» With the vision of **Trustworthy AI for Everyone**, the Strategy will be implemented in phases until 2025 through three sub-strategies that are technological, institutional, and ethical in nature, along with ten implementation tasks

| Vision | Trustworthy AI for Everyone | | |
|---|---|---|---|
| **Strategies** | **Create an environment for trustworthy AI** | **Lay the foundation For safe use of AI** | **Spread AI ethics Across society** |
| | ① Put in place a systematic process for securing trust for AI products/services | ① Make AI learning data more trustworthy | ① Provide strengthened education programs on AI ethics |
| | ② Support players in the private sector with securing trust for AI | ② Promote securing trust for high-risk AI | ② **Create and distribute checklists for each stakeholder** |
| | ③ Developing source technology for trustworthy AI | ③ Conduct assessment on influence of AI | ③ Operate a platform for ethic policies |
| | | ④ Improve regulations for increased trust for AI | |

3

## 2. Development of the Self-Checklist on National Guidelines for AI Ethics

> Since AI ethics is a relatively new social norm, **there is a gap that makes it difficult for individuals/organizations to apply unfamiliar AI ethics principles** in daily life or within organizations

» **As such, a self-checklist is needed so that individuals/organizations can judge and evaluate, for themselves, compliance with the National Guidelines for AI Ethics**

\* During the process of designing the National Guidelines, experts from industry and academia pointed out the need to develop specific guidelines that can serve as references for real-world implementation.

» In particular, SMEs and startups have **limited capabilities to establish and adhere to their own AI ethics standards** due to time, manpower, and resource constraints. This, in turn, can hinder their development and adoption of AI

**The Self-Checklist aims to support the voluntary efforts of people in relevant fields to practice AI ethics standards in actual work settings**

❶ **The Self-Checklist is closely linked to the National Guidelines' key requirements**

❷ **Self-assessment questions with high levels of generality were devised so that AI professionals can better utilize the Self-Checklist as a reference**

4

## 2. Development of the Self-Checklist on National Guidelines for AI Ethics : Progress (1/2)

— **Established a collaboration system with an organization that is preparing practical measures to enhance the trustworthiness of AI**

☑ Set up principles to minimize confusion that may be experienced by recommended users of the Self-Checklist during their self-assessment of each practical measure

— **Established and operated a research task force to develop the Self-Checklist**

☑ Identify essential questions for the Self-Checklist

* ① <u>Major public AI ethics self-checklists</u> from both at home and overseas,
  as well as ② <u>cases related to the negative effects of AI</u>, were collected, reviewed, and analyzed.

☑ Questions were derived after examining the necessity of hierarchical differentiation among questions, redundancy, contradictions, and other aspects

— **Received in-depth legal and technical consultations**

↓

**Prepared the first draft of the Self-Checklist as a reference to gather opinions**

## 2. Development of the Self-Checklist on National Guidelines for AI Ethics : Progress (2/2)

### Gathered opinions from varying stakeholders on the Self-Checklist ('21.10)

① **AI Ethics Policy Expert Council** (held twice)

» The Self-Checklist was revised to reflect the opinions of **10 experts** in AI-related technology, laws & institutions, and ethics policy.

② **Online information session for the Self-Checklist** (held twice)

» **34 AI experts** from academia, industry, the legal community, and civic organizations shared their ideas, and these ideas were reflected in the Self-Checklist.

### Released the first draft of the Self-Checklist and collecting public opinions (in progress)

☑ **Hosted an open policy seminar to enhance the trustworthiness of AI**

» The Self-Checklist was released to the public, and discussions among diverse stakeholders (e.g., companies and citizens) took place.

» Currently, public opinion is being heard through online channels.



## To release the final draft of the Self-Checklist to Practice AI Ethics Standards (planned)

# 3. Overview

**Self-Checklist on National Guidelines for AI Ethics (draft)**

### Purpose Of Use

**Realizing the three basic principles and ten key requirements of the National Guidelines for AI ethics**

The Self-Checklist supports voluntary efforts by stakeholders to enhance social acceptance and trust in AI technology and services

### Target Users

**Individuals/organizations that make decisions for the development/operation of AI systems**

In both public and private sectors, individuals or organizations in charge of establishing the purposes, procedures, and rules related to the development/operation of an AI system can utilize the Self-Checklist when they develop internal guidelines/rules that implement the National Guidelines by using the Self-Checklist as a reference and reflecting its content in their decision-making processes.

### How to Use

**As a reference for developing internal AI ethics guidelines**

Individuals and organizations in charge of designing/producing an AI system, implementing it through data and algorithms, or operating/managing it can practice the key requirements of the National Guidelines in actual work settings by following internal rules that reflect the content of the Self-Checklist

# 4. General Structure and Highlights

**① Structure of the Self-Checklist on National Guidelines for AI Ethics (draft)**

2021

The
Self-Checklist on
National Guidelines
for
AI Ethics (draft)

Ministry of Science and ICT     KISDI KOREA INFORMATION SOCIETY DEVELOPMENT INSTITUTE

## Contents

▸ **Background**

▸ **Purpose of Use and Target Users**

▸ **Table of Self-Checklist Questions**

▸ **Details on Questions Corresponding to Each of the Ten Key Requirements**

▸ **Appendix**

# 4. General Structure and Highlights

**② The Self-Checklist on National Guidelines for AI Ethics (draft)**

## Self-Checklist on National Guidelines for AI Ethics (draft)

· This self-checklist is recommended for individuals or organizations that make decisions for Artificial Intelligence (AI) system development and operation.

· Any individuals or organizations that establish a purpose and procedure and set regulations for AI system development and operation, in either the public or private sectors, can refer to and incorporate the items in this self-checklist into their decision-making process on internal guidelines or regulations for following the National Guidelines for AI Ethics.

· Any members of organizations or groups responsible for the design, configuration, implementation using datasets and algorithms, or maintenance and management of AI systems can follow internal guidelines with this self-check incorporated and practice the key requirements set forth in the the National Guidelines for AI Ethics while carrying out their tasks.

### E01. Human Rights

| | | YES | NO | N/A |
|---|---|---|---|---|
| E01.01 | Is the AI system developed and operated not to infringe on the human rights to life and safety? | | | |
| E01.02 | Is the AI system developed and operated to treat all humans as equal and without discrimination on any grounds, including gender, age, location, religion, race, ethnicity, economic status, and sexual orientation? | | | |
| E01.03 | Is the AI system developed and operated for its automated decision-making not to interfere with users' autonomous actions and decisions? | | | |
| E01.04 | Is the AI system developed and operated not to infringe on users' freedom of expression, the press, publication, assembly, and association? | | | |
| E01.05 | Is the AI system developed and operated to treat humans with dignity, such as not making users feel humiliated or ashamed? | | | |

### E02. Protection of Privacy

| | | YES | NO | N/A |
|---|---|---|---|---|
| E02.01 | In the case of collecting and using personal information for AI system development, did you carry out the self-check following the Artificial Intelligence (AI) Personal Information Protection Self-Checklist of the Personal Information Protection Commission in order to comply with the Personal Information Protection Act and other relevant laws? | | | |
| E02.02 | Do you trace the use of personal information during AI system development and operation? | | | |
| E02.03 | Did you assess the possibility of personal information leakage caused by reverse inference of pseudonymized or anonymized information used in AI system development and operation? | | | |
| E02.04 | Did you check whether the AI system created or used personal information based on the data collected online without the consent of data subjects? | | | |
| E02.05 | Did you establish any internal procedures to stop using the AI system if it would infringe on privacy or illegally collects and uses personal information? | | | |

### E03. Respect for Diversity

| | | YES | NO | N/A |
|---|---|---|---|---|
| E03.01 | Do you develop and operate the AI system not to unreasonably restrict the usage access of the socially disadvantaged? | | | |
| E03.02 | Did you make efforts to objectively determine and minimize the possibility of bias in the data utilized in AI system development in terms of gender, race, ethnicity, and nationality through a dedicated in-house team or an external expert or institution? | | | |
| E03.03 | Did you establish a set of procedures to listen to, review, evaluate, and reflect the opinions of stakeholders from a range of related fields in the AI system operation phase? | | | |
| E03.04 | In response to a case where bias, discrimination, or alienation would be found or suspected in the use of the AI system, did you establish a set of procedures to allow both developers and users to raise their objection or complaint internally or to the development organization and operator of the system and to review, assess, and reflect on the issue internally? | | | |
| E03.05 | Do you provide or utilize educational programs to help AI system developers improve their perception and analysis ability on possible biases in the AI system? | | | |

### E04. Prevention of Harm

| | | YES | NO | N/A |
|---|---|---|---|---|
| E04.01 | Did you review in advance and take preventive measures against the possibility of whether the AI system would pose a threat to human lives or cause physical harm to humans or their property? | | | |
| E04.02 | If you discovered the possibility that the AI system could pose a threat to human lives or cause physical harm to humans or their property while being used beyond its original purpose, did you notify users of such information in advance? | | | |
| E04.03 | Did you establish a procedure for users to file a claim and present their opinion for any unexpected damage incurred during their use of the AI system? | | | |
| E04.04 | Did you establish internal procedures to notify relevant agencies of unexpected damage or harm discovered or reported during the application of the AI system and stop using it to prevent the further spread of such damage? | | | |

### E05. Public Good

| | | YES | NO | N/A |
|---|---|---|---|---|
| E05.01 | Did you consider the possibility that the AI system would damage the public interest or cause dysfunction by representing the interests of only particular individuals or groups? | | | |
| E05.02 | Did you consider the possibility of side effects from the use of the AI system, such as stimulating violence, obscenity, or a speculative spirit? | | | |
| E05.03 | Did you consult with external experts when it was difficult for internal developers or operators to review and reach conclusions on the positive or negative social and economic impacts of the AI system? | | | |

9

# 4. General Structure and Highlights

**② The Self-Checklist on National Guidelines for AI Ethics (draft)**

## E06. Solidarity

| | | YES | NO | N/A |
|---|---|---|---|---|
| E06.01 | In the AI system development stage, do you monitor trends in the development of the international standards for ethical AI discussed at ISO/IEC JTC1/SC42? | | | |
| E06.02 | Do you provide opportunities for users from all walks of life to communicate or interact with each other under the purpose of AI system development and operation? | | | |
| E06.03 | Did you consider the possibility of whether the use of the AI system would hinder social cohesion, such as by causing conflicts between different communities, genders, generations, and classes? | | | |
| E06.04 | Did you consider the use of low-carbon energy sources during AI system development and operation to participate in the international community's efforts for carbon neutrality? | | | |

## E07. Data Management

| | | YES | NO | N/A |
|---|---|---|---|---|
| E07.01 | Do you evaluate objectively and periodically whether the data used in AI system development are of high quality and optimal through a dedicated internal department or external expert or institution? | | | |
| E07.02 | Do you track the sources and main processing processes of the data used in AI system development? | | | |
| E07.03 | Do you use appropriate performance indicators to draw correct statistical conclusions (type I error*, type II error**, etc.) when evaluating whether the AI system uses quality data and performs accurate and valuable predictions?  * Type 1 error: a false positive conclusion, ** Type 2 error: a false negative conclusion | | | |
| E07.04 | Is there any mechanism established for the supervision of the data collection and processing used in AI system development, and are only authorized users allowed to access the data analysis and management under control? | | | |

## E08. Accountability

| | | YES | NO | N/A |
|---|---|---|---|---|
| E08.01 | Did you check potential risks that may occur during AI system development and operation, including mental or physical harm to humans, damage to property, or adverse social effects, and develop countermeasures (i.e., introducing technology for risk minimization, benchmarking best practices, and performing periodic review)? | | | |
| E08.02 | Do the AI system developer and development team appropriately archive key information on the flow of algorithms and the decision-making processes? | | | |
| E08.03 | Do the AI system developer and development team make efforts to improve the following abilities?  – Ability to judge the suitability and risks of the fields where the AI system is applied to prevent possible damage  – Ability to detect and manage bias inherent in data  – Ability to interpret the output and decisions of the AI system  – Knowledge of to whom and through which procedure to report any potential damage perceived during the use of the AI system | | | |
| E08.04 | In order to comply with AI ethics and ensure accountability during AI development and operation, did you create an internal organization and define the roles of its participants? | | | |
| E08.05 | Are the roles and responsibilities appropriately defined for different types of stakeholders, organizations, and members of organizations in order to clearly assign accountability for damage and losses incurred by AI system development and operation?  – Did you establish procedures for dispute mediation in response to cases where it is unclear who should be held accountable for damage and losses. | | | |
| E08.06 | In the case of an AI system that has a significant mental and physical impact on humans or their property, did you establish procedures to discontinue or recall the system on the market, report to the government and competent authorities, and notify users if accountability can no longer be sustained due to issues in AI system development and operation? | | | |
| E08.07 | Did you prepare a reserve plan (such as signing up for an insurance policy or fund or preparing a reserve) for providing reasonable compensation and rewards for damage incurred by the use of the AI system? | | | |

## E09. Safety

| | | YES | NO | N/A |
|---|---|---|---|---|
| E09.01 | Do you operate a safety plan or measures to respond to significant risks (i.e., system error, hacking, malware installation, adversarial attack) that may occur during AI system development and operation? | | | |
| E09.02 | Did you develop measures such as monitoring advanced technologies to mitigate or confront purposeful attacks on the AI system? | | | |
| E09.03 | Did you establish safety measures (i.e., emergency stop function, user notification, and system closure) in response to risk factors or accidents that can be caused by the automated decision-making of the AI system? | | | |
| E09.04 | Did you prepare safety measures against abnormal operation or unexpected errors of the AI system? Do you provide shareholders and users with sufficient information on the effective implementation of these measures and their limitations? | | | |
| E09.05 | Did you evaluate in advance the risks that may arise from human interaction with the AI system (i.e., inappropriate attachment, addiction, human monitoring by AI) and prepare measures to mitigate them? | | | |
| E09.06 | In order to continuously evaluate the safety of the impacts of the AI system, did you develop measures such as the reflection of user feedback and periodic expert evaluation through an internal team or external agency? | | | |

## E10. Transparency

| | | YES | NO | N/A |
|---|---|---|---|---|
| E10.01 | During AI system development, did you design it to allow managers or users and other stakeholders to interpret important decisions related to humans?  – Specifically, do you transparently disclose the following information for each type of decision-making or field of application: source of training data, data collection and processing method, data movement path, data update cycle, major factors influencing the results, model change, and subject of change | | | |
| E10.02 | Do you notify users of the fact that the application products and services of the AI system make decisions based on AI algorithms and that users are interacting with AI? | | | |
| E10.03 | Do you provide guidebooks and manuals for users and other stakeholders in order to ensure that the AI system is used for its intended purpose? | | | |
| E10.04 | When users ask for explanations about the data collected by the AI system, the main factors affecting the impacts of decision-making, etc., can you answer such requests with understandable explanations? | | | |

# 4. General Structure and Highlights

③ Examples Highlighting the Structure of the Self-Checklist (draft)

---

**Requirement 3 Respect for Diversity**

- All stages of AI development and utilization should reflect the diversity and representativeness of its users. Bias and discrimination based on personal characteristics such as gender, age, disability, ethnicity, religion, and country should be minimized. Commercialized AI systems should be used in ways that treat all individuals in an impartial and equal manner.
- Access to AI technologies and services by socially disadvantaged or vulnerable groups should be ensured. There should also be efforts to distribute the benefits brought about by AI to people from all walks of life rather that to specific subgroups.

**Checklist**

**E03.01** Do you develop and operate the AI system not to unreasonably restrict the usage access of the socially disadvantaged?

**E03.02** Did you make efforts to objectively determine and minimize the possibility of bias in the data utilized in AI system development in terms of gender, race, ethnicity, and nationality through a dedicated in-house team or an external expert or institution?

**E03.03** Did you establish a set of procedures to listen to, review, evaluate, and reflect the opinions of stakeholders from a range of related fields in the AI system operation phase?

**E03.04** In response to a case where bias, discrimination, or alienation would be found or suspected in the use of the AI system, did you establish a set of procedures to allow both developers and users to raise their objection or complaint internally or to the development organization and operator of the system and to review, assess, and reflect on the issue internally?

**E03.05** Do you provide or utilize educational programs to help AI system developers improve their perception and analysis ability on possible biases in the AI system?

**Reason for Proposal**

AI systems can follow and reproduce the stereotypes and prejudices inherent in society, which results in bias, suitability, and equity issues. This is attributable to the fact that human biases can be reflected in the data used for AI system development and operation. Efforts to minimize discrimination in the application of AI systems include not only reducing biases inherent in data or guaranteeing access to technology or services, but also engaging stakeholders from different socio-economic backgrounds from the development and operation stages. Therefore, efforts to consider various members of society and to prepare evaluations and procedures to minimize biases and discrimination are assessed.

**REF1 Digital Dubai 'AI ETHICS GUIDELINES*(2020)'**

- Digital Dubai's AI Ethics Self-Assessment Tool describes the value of respect for diversity in the context of fairness and suggests examining whether AI systems can be accessed and used fairly by different user groups and assessing them with the following items.

"Efforts could be made to include people from diverse demographic backgrounds in the development and deployment processes."

"AI systems informing significant decisions should be developed by diverse teams with appropriate backgrounds."

* Digital Dubai, 'AI ETHICS GUIDELINES'(2020) (https://www.digitaldubai.ae/initiatives/ai-principles-ethics)

- 14 -

---

**REF2 Google 'Artificial Intelligence at Google: Our Principles*(2018)'**

- It offers guidance on the ethical principles that the developers of AI applications at Google should consider.
- One of key principles is to avoid creating or reinforcing unfair biases, while seeking to avoid unjust impacts based on any characteristics such as race, ethnicity, gender, nationality, income, sexual orientation, ability, and political or religious belief.

* Google, 'Artificial Intelligence at Google: Our Principles' (2018) (https://ai.google/principles)

**Case Study Cases of Key Requirement 3 "Respect for Diversity"**

☞ If AI provides outcomes biased towards a specific race, ethnicity, gender, or age group and is used in recruiting or credit evaluation, this could lead to unequal opportunities for particular groups.

(Related study) More than 90% of individuals exposed to logging job postings were men, and more than 70% of them were white. Over 65% of cleaning job postings were shown to women, and over 75% of them were black. Since the research team posted the job openings under the same conditions, they inferred that the results were due to the algorithms of targeted advertising platforms (Ali et al., 2019).

(Case 1) In January 2021, AI chatbot Iruda was suspended after it used words of gender and racial discrimination and engaged in hate speech towards minorities such as people with disabilities and members of the LGBT community.



(Case 2) The Dutch government developed and applied the System Risk Indication (SyRI), a system designed to detect welfare benefits and tax fraud. However, a Dutch court ruled to halt the use of SyRI due to the possibility of its discriminatory application to minorities and vulnerable groups.

- 15 -

11

# 5. Future Plans

## 01
### Discover outstanding use cases

Support the development of internal guidelines that utilize questions from the Self-Checklist and discover exemplary use cases

## 02
### Develop implementation tools
**for the Self-Checklist**

Create online self-assessment tools to enhance convenience and access to the Self-Checklist

## 03
### Complement questions
**in the Self-Checklist**

After pilot distribution, continue to upgrade the Self-Checklist by reflecting user feedback

# Thank You