Korean Women's Development Institute ISSUE PAPER

KWDI Issue Paper

Publisher: Korean Women's Development Institute
Publication Date: July 15, 2023

Research Title: Gender Bias in the Use of Artificial Intelligence Deep Learning and Suggestions for Improvement
Principal Researcher: Moon Meekyung, Research Fellow

Policy Efforts Needed in Response to the Increase in Artificial Intelligence Gender Bias

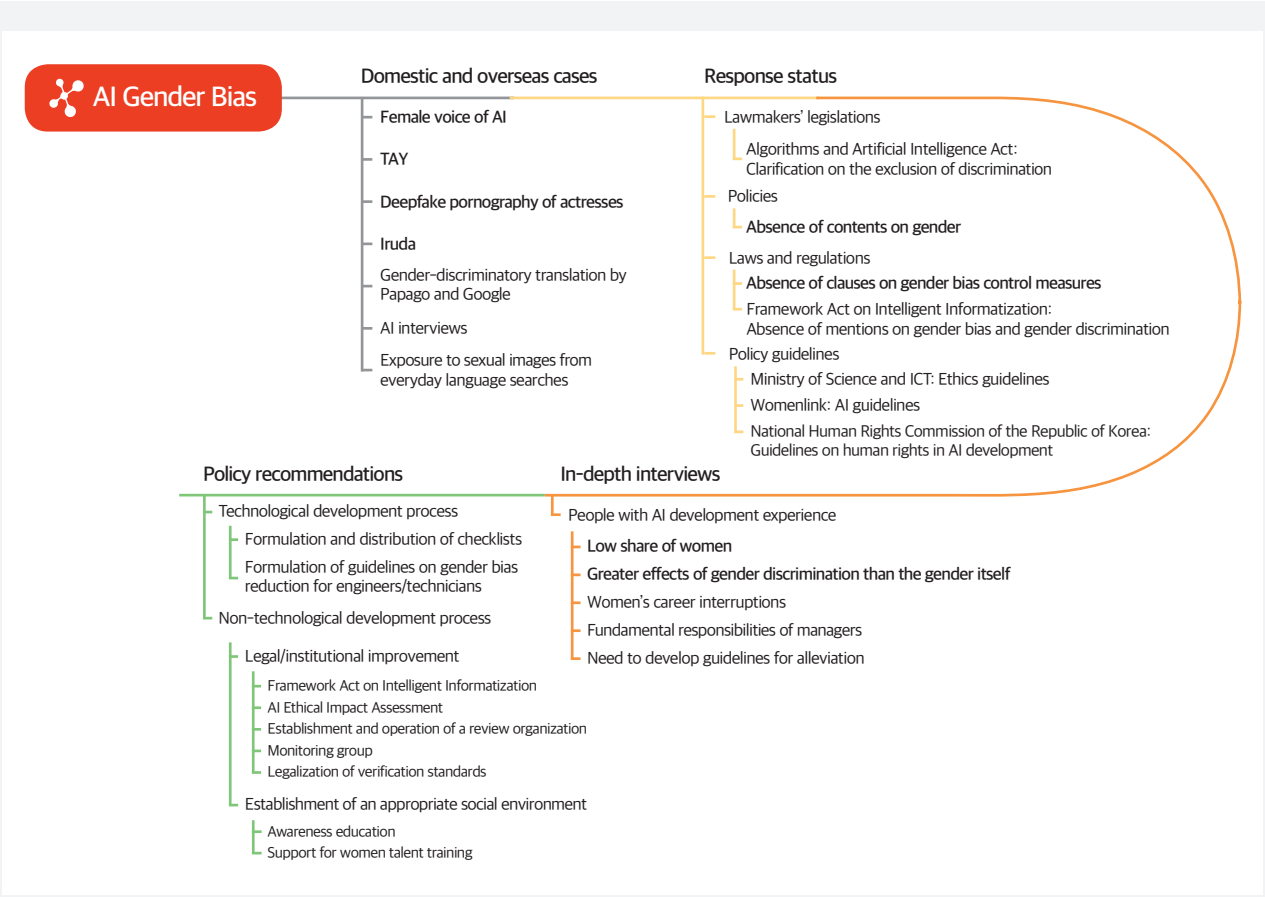
Abstract

- Society is witnessing its inherent gender discrimination through algorithms, as seen in the learning and production of hate speech by artificial intelligence (AI) chatbots "Iruda" and "Tay".
  - In addition to the events that captured the media attention, problems have been identified in AI technologies used in everyday life, such as sexist remarks produced by AI speakers, gender-discriminatory translation functions, and discriminatory evaluations made by AI job interviews, and so on.
- Thus, this study examined gender biases that occur in the use of AI deep learning and explored the current efforts in addressing the biases. The study also proposed policy measures based on in-depth interviews and expert opinions.
  - This study categorized domestic and overseas AI gender bias cases and diagnosed the causes, and thereby found that two aspects of responses should be considered together to alleviate AI gender bias.
  - First, technical guidelines are needed in order to consider and reduce AI gender bias at each stage of the technical engineering level ranging from technology configuration to utilization. Appropriate guidelines are proposed in this study.
  - Second, gender stereotypes and gender discrimination in society are the ultimate sources of AI gender bias. Thus, it is vital to recognize that gender bias in AI technologies is a problem for the whole society. Acknowledging this, the study proposes policies for improving relevant laws and institutions and also for building a social environment to reduce gender bias in society as a whole.

KWDI Issue Paper

Please cite this Issue Paper as follows.

Moon, M., Kim, B., Kang, K., Kim, Y., Kim, H. and Heo, Y. (2022). Gender Bias in the Use of the Artificial Intelligence Deep Learning and Suggestions for Improvement. Seoul: Korean Women's Development Institute.



Background and issues

- With the rapid development of AI technology in recent years, great attention has been on AI's social and economic aspects. People expect that AI's decision-making, unlike human decision-making, would be neutral and free of bias or prejudice, as AI decision-making is a machine process. However, given the current level of technology, AI is closer to a statistical analysis tool that derives the best possible answer under given constraints than an independent decision-making agent.

- The society has witnessed cases such as the collective sexual violence involving AI chatbot "Iruda", the controversies over hate speeches produced through sexual harassment learning, and the AI chatbot "Tay" that learned and repeatedly produced hate speech on its own. AI that has learned specific discriminations or their mechanisms can be used for credit transactions and loans, evaluation of job candidates, admissions evaluation for educational institutions such as universities, selection and provision of personalized articles, or election and recommendation of persons for other specified purposes (Heo, 2018). These examples of AI algorithms not only illustrate that unfair discriminations can have consequences and effects, but also show that AI-generated discriminations reflect the discriminations that are inherent within society, such as those involving race and gender, and can even further reinforce those discriminations.

To reduce gender bias in AI, it is essential to incorporate an approach on addressing the issue within a cyclical link between humans and the environment, and this includes addressing people's biased perceptions on gender and the social environment that affects the production of such perceptions. Given the rapid development of AI deep learning technology and its widespread social impact, this research analyzed the status of gender bias in the AI development process and sought ways to alleviate it. To this end, the study analyzed examples of biases caused by AI with a focus on gender, and examined the current status on laws and policies at home and abroad that are designed to address biases. Based on the implications derived from the analyses, the study discussed ways to alleviate gender bias through laws and institutions, measures necessary to build an appropriate social environment, and guidelines through an engineering approach that can be understood and applied in the AI technological development.

Investigation and analysis results

Gender bias in AI

- AI is a collection of concepts and technologies that offer different meanings to people, such as autonomous cars, robots mimicking humans, machine learning, and so on, and AI's application programs are ubiquitous. The recent definition of AI states that AI is a field of computer science and information technology that studies the ways computers perform sentence understanding, video recognition, voice recognition, and learning that can be done through human intelligence, thereby allowing computers to mimic human intelligent behavior. Through AI, computers can utilize vast amounts of data and use learned intelligence to produce results in much less time than the time required by humans.

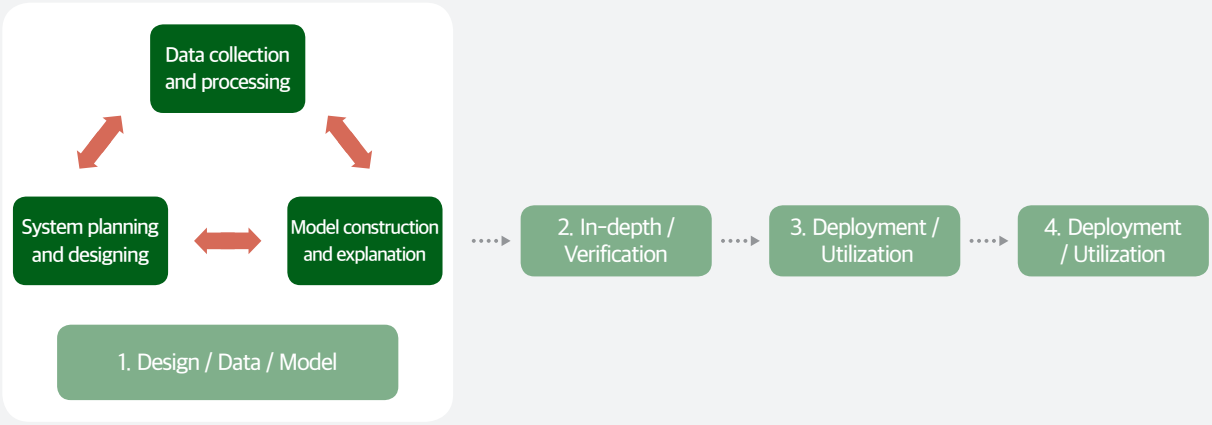
- Gender bias in AI tends to arise largely from biased algorithm design, biased data (collection, processing, exposure, etc.), and baseless expectations on AI's objectivity (which assumes that machines are less biased than humans). Addressing these problems is not easy given the difficulties for humans to accurately understand the process and basis of AI outputs, and it is also challenging for the general public to easily access such information due to the characteristics of AI technology (including problems with transparency or feasibility of explanations).

Analysis of domestic and overseas AI gender bias cases

- This study investigated domestic and overseas cases of AI gender bias by technological development type. The production and utilization of AI technology systems are largely divided into several phases including: the planning of design, data, and modeling; system verification and validation; deployment for the practical use of the model as a service; overall operation; and subsequent monitoring.

- An analysis of the domestic and overseas cases revealed that the phases were categorized into "AI planning and design phase", "data processing phase", and "modeling phase, including algorithm generation and learning".

<Figure 1> AI technology configuration phases



Source: OECD (2019; as cited in translation by National Information Society Agency, 2019)

<Table 1> Examples of gender bias by AI technological development type

AI planning and designing phase	
Gendered AI	Female voice and images of AI secretaries, social robots, etc., and adaptive responses to sexual harassment, etc.
Fusion between sexual exploitation, profit-making, and AI technologies	Pornography using deepfake technology Development of algorithms to track women's identities, etc.
Gender target research	Research on facial recognition algorithms that can identify sexual orientation with facial images only
Data processing phase	
Bias in data itself	Automatic translation and linking of gender-neutral or feminine words to masculine words during machine translation and word embedding in language processing Automatic data labeling or automatic image generation algorithms producing gender-specific results
Gender bias in setting the sample group for data collection	Gaps in recognition rates between white men and black women in facial recognition programs Male group bias in medical data
Modeling phase, including algorithm generation and learning	
Algorithm designs without any consideration of negative gender bias or its improvement	Gender gaps in the advertisement exposure for career development in Science, Technology, Engineering, and Math (STEM) Computer perspective model training that reinforces traditional gender stereotyping
Algorithm machine learning based on gender-biased data	Gender gaps in the advertisement exposure for high wage jobs concentrated in men Unconditional deduction of points for women-related key words in the AI recruitment process
Difficulty of ensuring algorithm transparency or feasibility of explanation	Low credit limits set for women whose assets are in joint name with their husband Search overexposure of suggestive contents when searching key words related to women of colored race by presenting the search results on the top, etc.

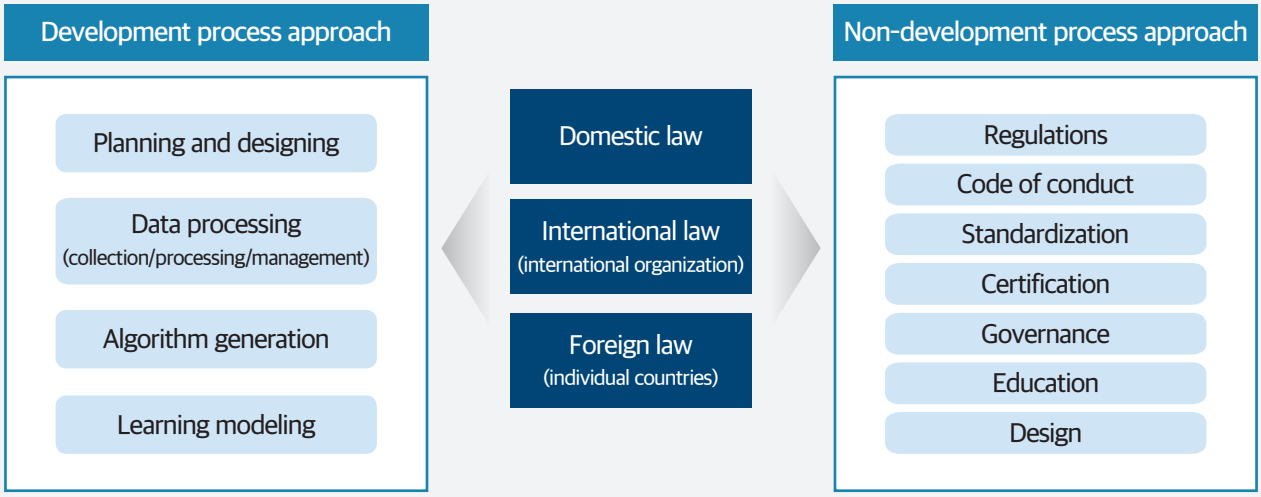
1) The technological development process entails a technical approach in which AI is planned, developed, and produced, whereas the non-technological development process refers to an approach that involves AI-related human rights, ethical issues, and so on.

▶ The results showed that gender bias occurred throughout all phases of AI technology configuration. Thus, useful tools need to be developed to consider AI gender bias issues and reduce gender bias at each phase. The results also suggested that the fundamental reasons for the difficulties in developing and sustaining practical measures included the following: the accumulation and impact of historical and structural gender biases; the technology industry that fails to recognize the issue as a major challenge for society; and a lack of awareness among the general public.

Current status on domestic and overseas responses to AI gender bias

▶ This study analyzed current measures at home and abroad as well as international organizations by categorizing AI-related laws and policies into technological development process and non-technological development process.

<Figure 2> Current status on responses to AI gender bias for technological development process and non-technological development process



▶ Current AI-related domestic legislations and policies recommend that discrimination and bias should not occur at any stage of AI technology, but most legislations lack provisions that explicitly consider gender.

<Table 2> Status of domestic responses

Laws	
Technological development process	There is no provision on measures to prevent gender discrimination or control gender bias.
Non-technological development process	Although the "Framework Act on Intelligent Informatization (Law No. 17344)" mentions concerns about inequality or gaps in the broad sense, there is no explicit statement on gender bias or gender discrimination.
Policy	
Technological development process	Despite the "Artificial intelligence (AI) R&D strategy for realizing I-Korea 4.0", there is an insufficient policy approach to detailed development process, including data processing, algorithm generation, and learning modeling, and there is no consideration of gender elements in the overall policy.
Non-technological development process	There is no gender-related content.
Bills proposed by lawmakers	
Technological development process	The "Bill on Algorithm and Artificial Intelligence" clearly stipulates the exclusion of discrimination for the reasons of gender, etc.
Non-technological development process	Although the "Bill on Artificial Intelligence Research and Development, Industrial Promotion, Ethical Responsibility, etc." offers an explicit provision on the protection of human rights and dignity, there is no perspective on gender equality.
Policy guidelines	
Technological development process	"Ethics Guidelines for an Intelligence Information Society" by the Ministry of Science and ICT "Human Rights Guidelines for the Development and Use of AI" by the National Human Rights Commission of the Republic of Korea "AI Guidelines Made Together by Feminists" by the Womenlink
Non-technological development process	"Human Rights Guidelines for the Development and Use of AI" by the National Human Rights Commission of the Republic of Korea "AI Guidelines Made Together by Feminists" by the Womenlink

- International organizations such as the Organization for Economic Co-operation and Development (OECD), the European Union (EU), and United Nations Educational, Scientific and Cultural Organization (UNESCO) have proposed recommendations on AI ethics, emphasizing that each country should ensure reliability and consider AI ethics in developing and utilizing AI technology. As seen in the cases of legislations by major countries, it is vital to include gender bias when transparently evaluating and disclosing the ripple effects of AI on racial and gender discriminations.

- It is critical to incorporate contents that address gender bias reduction in the Framework Act on Intelligent Informatization which forms the basis of AI technology in South Korea. Considering the AI ethics recommended by UNESCO, it is also essential to include contents on gender bias reduction when establishing and operating dedicated organizations that are tasked with assessing and monitoring ethical impacts.

In-depth interviews

- In-depth interviews were conducted with 10 people (six men and four women) who had AI development experience. The interviews consisted of questionnaire surveys (phase 1), in-depth interviews on socio-cultural environment (phase 2), and surveys regarding the competency and awareness on bias reduction techniques (phase 3).
- The results showed that the main reason for AI gender bias was insufficient attention at the data collection, selection, and processing stages. Fundamental responsibility lies with the final decision makers (managers) who set directions for AI configuration. However, it is difficult to apply bias mitigation techniques and elements due to the limitations of time and funding in the development processes, thus it is imperative to develop guidelines for reducing gender bias at the technological level.
- The results confirmed that currently the proportion of women was remarkably low among the developers in their 30s in the AI R&D field, and the interviewees were aware that the experience of gender discrimination rather than gender itself tended to affect the development processes. In particular, it was acknowledged that a sexist work environment in the cutting-edge science and technology field could work as a major factor in interrupting women's careers.
- Through the in-depth interviews, this study proposed recommendations on: 1) gender bias checklists for AI ethics; 2) guidelines for mitigating overall bias in the AI technological development phase for engineers; 3) guidelines for mitigating gender bias by AI technology for developers and suggestions for building an appropriate social environment, including gender balanced human resources development and so forth.

Policy recommendations

According to the categorization of domestics and overseas AI gender bias cases and the assessment of the causes, two aspects of responses are required to reduce AI gender bias.

- 1. Responses should aim to mitigate gender bias at the technological and engineering levels in all stages of technological configuration and utilization. As confirmed earlier, AI gender bias can occur in each phase of AI technology configuration. Therefore, it is suggested that the issue of AI gender bias be taken into account throughout the entire process of modeling AI technology configuration, including AI planning and design, data processing (collection, processing, management, etc.), and algorithm generation and learning, etc. and that technical guidelines be prepared to reduce the bias in each stage.
- 2. The ultimate sources of AI gender bias include gender stereotypes and discrimination in the society. Therefore, it is suggested that gender bias in AI technologies be recognized as an issue of the whole society, and that laws and institutions be improved by incorporating the implications from the AI-related legislations and policies of international organizations. This study also recommends that relevant policies be established to create a social environment that is conducive to reducing gender bias in the general public.

<Table 3> Policy suggestions by development process type

1. Improvement measures for the technological development process	AI technology and system set up stage
	(1) Develop and disseminate gender bias checklists for AI ethics (2) Develop and disseminate guidelines for engineers to reduce gender bias in the AI technological development process (3) Develop and disseminate guidelines for developers to reduce gender bias for each AI technology
2. Improvement measures for the non-technological development process	Improvement in laws and institutions
	(1) Explicitly state the consideration of gender in the basic principles of the Framework Act on Intelligent Informatization (Article 3) (2) Explicitly state gender impact assessment in the Framework Act on Intelligent Informatization (Article 56) (3) Include a clause on the prohibition of gender bias in the AI Ethical Impact Assessment (4) Organize and operate an AI ethics review body from the gender sensitive perspective (5) Operate an AI gender bias monitoring team at the government level (6) Legislate verification standards for data- and algorithm-related gender bias
	Build an appropriate social environment
	(1) Provide education on AI ethics and gender bias (2) Support human resources training and career development for women in AI technology and basic science

References

Naver Encyclopedia - Artificial Intelligence

(https://terms.naver.com/entry.naver?docId=1136027&cid=40942&categoryId=32845, Accessed on April 12, 2022)

Algorithms of the Brain and AI

(https://misterio.tistory.com/entry/%EB%87%8C%EC%99%80-%EC%9D%B8%EA%B3%B5%EC%A7%80%EB%8A%A5AI%EC%9D%98-%EC%95%8C%EA%B3%A0%EB%A6%AC%EC%A6%98Algorithm, Accessed on March 17, 2022)

Yang, J. (2017). Effects of Bias and Opacity of Artificial Intelligence Algorithms on Legal Decision Making and its Discipline. Korean Lawyers Association Journal, 2017-6.

Heo, E. (2018). The Preliminary Consideration for Discrimination by AI and the Responsibility Problem - On Algorithm Bias Learning and Human Agent -. Korean Feminist Philosophy 29.

OECD (2019:15). Artificial Intelligence in Society. Paris: OECD Publishing.

Responsible Ministries	:	Artificial Intelligence Policy Division, Ministry of Science and ICT; ICT Policy Coordination Division, Ministry of Science and ICT; Women's Human Resources Development Division, Ministry of Gender Equality and Family; Rights and Interests Infringement Prevention Division, Ministry of Gender Equality and Family
Relevant Ministries	:	Artificial Intelligence Policy Division, Ministry of Science and ICT; ICT Policy Coordination Division, Ministry of Science and ICT; Women's Human Resources Development Division, Ministry of Gender Equality and Family; Rights and Interests Infringement Prevention Division, Ministry of Gender Equality and Family