2023
Summary of Research Papers 03

Gender Bias in the Use of Artificial Intelligence Deep Learning and Suggestions for Improvement

Meekyung Moon
Bok-tae Kim
Kyung-ju Kang
Yesol Kim
EuSun Heo
HyoEun Kim

Korean Women's Development Institute

Gender Bias in the Use of Artificial Intelligence Deep Learning and Suggestions for Improvement

Korean Women's Development Institute
225 Jinheung-ro Eunpyeong-gu
Seoul, 03367, Republic of Korea
www.kwdi.re.kr

Contents

Tables

Figures

Gender Bias in the Use of Artificial Intelligence Deep Learning and Suggestions for Improvement

Meekyung Moon
Bok-tae Kim
Kyung-ju Kang
Yesol Kim
EuSun Heo
HyoEun Kim

Ⅰ. Introduction

○ There is always the possibility of artificial intelligence (AI) algorithms causing bias or errors. In particular, bias in AI is likely to lead to discrimination and to significantly undermine the social value of anti-discrimination (Yang Jong-mo, 2017:64; Byeon Soon-yong, 2020:146 recited[1]). Because AI is created by humans who tend to be biased and oriented towards evaluations, it cannot be guaranteed that AI is always fair and neutral. Therefore, against the backdrop of AI increasingly replacing humans via automated technologies, AI-induced discrimination should be recognized as a key social issue and discussed in depth.

○ Most literature regarding AI-driven discrimination has focused on

[1] Byeon Soon-yong (2020). Research on Bias in AI in the Context of Data Ethics. Vol. 128, Journal of Ethics

discussions about generally mitigating bias in AI and improving relevant situations. There has been little literature that addresses and highlights the issue of gender bias. Key tech firms tend to concentrate on diversity and inclusiveness, rather than gender bias and discrimination. Moreover, such endeavors are currently deemed to be superficial. As a result, demands for technology assessment and relevant legislation have grown, but currently it is simply under discussion.

○ Gender bias in AI is not limited to technical issues. Considering cyclical connections between humanity and circumstances such as human's biased recognition of gender and social environments affecting such attitudes, improvement plans should be devised. Therefore, it is necessary to prepare bias mitigation plans for reflecting gender perspectives in legal and institutional contexts, enhance administrators' (AI operators and planners) and program developers' recognition of gender bias and provide gender bias-related guidelines in terms of systems and engineering that are easily available to them.

○ Focusing on the rapid development of AI and deep learning technologies and on their impact on society, this study is designed to analyze and reduce gender bias seen in the process of developing AI. This paper also aims to seek ways to mitigate gender bias in the context of legal and institutional systems and social environments.

Ⅱ. Gender Bias in Artificial Intelligence (AI)

1. Concept of Gender Bias[2)]

○ The recent introduction of AI has led to the issue of bias in AI being brought into the spotlight. In particular, it seems to be clearly associated with racial and gender issues, highlighting social concerns and ethical, social, legal, and technological endeavors to seek solutions thereto. In connection therewith, algorithms examining gender bias are also under development. Systems for the ethical verification of algorithms are expected to be introduced in the future.

○ However, it is the concept of gender bias that should be first examined when discussing gender bias in AI. Gender bias refers to biased impact by gender or on gender witnessed in the process of developing technologies or using AI such as AI system planning/ design, data processing (collection, processing, and management), and algorithm generation and learning (modeling) that belong to the initial stage of AI technology development. In this context, bias refers to ethical bias creating or strengthening moral prejudice, rather than statistical bias. This is more true when addressing gender bias in AI. This needs to be classified into ‘gender bias in data and algorithms in a logical context’ and ‘gender bias in AI and in the use of AI’ seen in the phase of AI utilization. Herein, gender bias is defined not only as bias in data, algorithms, and utilization but also as gender discrimination or gender-biased phenomena witnessed in the process of the aforementioned phases being mixed.

---

2) Based on data given by professor Byeon Soon-yong (Seoul National University of Education) at this research advisory meeting (Oct. 13, 2022)

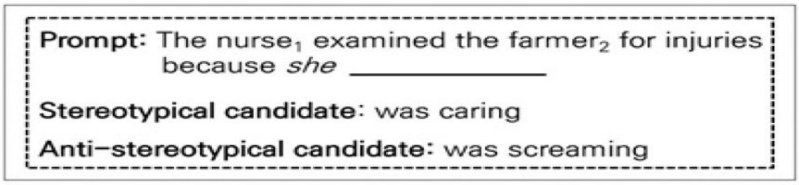
○ This paper focuses on bias in data, algorithms, and utilization (gender bias in AI and in the use of AI), subsets of gender bias.

2. Gender Bias in Data

○ Gender bias in data occurs when gender bias in data themselves becomes an issue in the process of generating AI learning data. In cases where unrefined learning data are used, AI learns racial, age, and gender discrimination without modifications or cultures and values belonging to specific languages, raising the issue of diversity or fairness. For instance, GPT-3 learns Common Crawl based on a huge amount of web data accumulated over eight years. However, most of the web data are texts written by white males in their 20s to 30s, native speakers of American and British English. When it was disclosed that GPT-3 data were created by white American and British men, racial and gender bias in learning data emerged as a key issue. This means that AI exposed to potential gender bias in data is likely to use gender-discriminatory expressions.

3. Gender Bias in Algorithms

○ It is impossible for flawlessly gender-neutral data to exist in the process of learning algorithms. Therefore, given that being exposed to gender bias to a certain degree cannot be avoided, the process of algorithms being designed or utilized and the resulting data will naturally show bias. As shown in the figure below, algorithms for assessing or verifying gender bias in AI have been developed. In the process of uncovering the gender of the nurse or the farmer using possible answers for the blank line, gender bias is deemed to be assessed and measured.



Source: Research Team

[Figure 1] Gender Bias Assessment Algorithm (Sample)

4. Gender Bias in AI

○ A significant set of examples of AI presents the following question: Does an AI assistant default to a female voice? Korea's AI chatbot Iruda has the identity of a female college student and female sex robots were released first, which can be understood as a form of gender bias in AI. Therefore, the fact that AI assistants or social robots are designed and released as feminine forms clearly shows gender bias in AI.

5. Gender Bias in the Use of AI

○ Gender bias in the use of AI is similar to gender bias in AI in the context of concept. The former refers to cases where people show biased attitudes toward AI in terms of gender. For example, in 2016, users led Microsoft's chatbot Tay to learn gender discrimination and thereby to use such types of expressions. Illegally using AI-based deep fakes or digital humans as a means for sexual exploitation and income creation can also be regarded as one such case. Gender bias in AI means bias included in AI itself while gender bias in the use of AI refers to bias that AI users display.

Ⅲ. Cases of Gender Bias in AI

1. Generation of Gender Bias in AI

○ Artificial intelligence used for self-driving vehicles, robots imitating humans, and machine learning is deemed to be a large group of concepts and technologies bringing diverse advantages to humanity. Relevant application programs can be seen everywhere. In terms of definition, AI is a subset of computer engineering and IT striving to enable computers to understand sentences, implement learning processes, and recognize videos and voices just like humans do. In other words, it is designed to make it possible for computers to imitate humans. AI enables computers to utilize a huge amount of data and learned intelligence in order to produce deliverables much faster than humans.

○ Gender bias in AI can be found in algorithm design and data (collection, processing, exposure, etc.) and in the groundless belief of AI's objectivity (assumption that machines are less biased than humans). However, it is not easy to react to such bias because it is difficult for humans to accurately understand the process of AI producing results, as well as relevant background, and to have access to such information due to the features of AI technologies (transparency or explainability).

2. Local and Global Cases of Gender Bias in AI

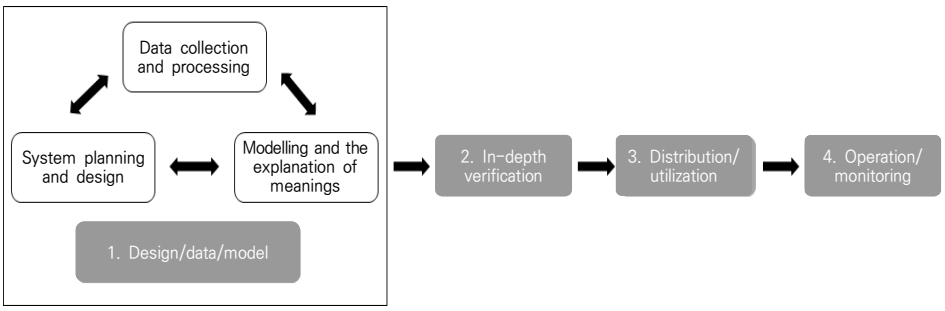
○ Cases were collected and selected by reviewing literature and research regarding AI ethics and bias such as AI principles and recommendations. To figure out trends in addressing the issues, AI

technologies used in modern society for the past decade were discussed, reviewing AI ethics-related literature concerning AI principles and bias and the impact of such technologies on society. Focusing on publications and papers released by actors (individuals or organizations) substantially impacting AI technologies and technology governance, cases where bias issues are mentioned were generally examined. Then, cases mentioned repeatedly and having significant effects on society were chosen, considering the diversity of actors such as governments (including the Korean government), international organizations, tech firms, and academia. The scope of case search expanded to include papers regarding causes and improvement plans, as well as local and global policy documents, research literature, and corporate data concerning follow-up and improvement measures.

○ Selected cases were assessed based on the concepts and causes of bias in AI, countermeasures, and special notes discussed in prior research. In accordance with research objectives, the issue of gender bias was especially highlighted, reclassifying cases by technology development phase, rather than simply listing issues, and categorizing cases based on what they have in common. As a result, seeking substantial responses by technology configuration phase, this study explained that gender bias in AI by technology development phase is not limited to technical issues.

○ The generation and utilization of AI technology systems are classified into ‘design, data, and modelling planning,’ ‘system verification and validation,’ ‘deployment for the real use of models,’ ‘operation,’ and ‘monitoring.’ Local and global cases were

categorized into ‘AI planning and design,’ ‘data processing,’ and ‘modelling for algorithm generation and learning.’



Source: OECD (The 2019 Translation Recited, 2019; National Information Society Agency)

[Figure 2] AI Technology Configuration Phase

〈Table 1〉 Cases of Gender Bias by Type of AI Technology Development

AI planning and design	
Gendering of AI	Female voices and imaging of AI assistants and social robots, accepting sexual harassment, etc.
Combination of AI technologies and sexual exploitation and income creation	Pornography based on deepfake technologies, development of algorithms for identifying women, etc.
Research on gender targets	Research on face recognition algorithms for identifying sexual orientation via face images
Data processing	
Bias in data themselves	Automatically translating or connecting gender-neutral or feminine words into and to masculine words in the process of machine translation and word embedding. Auto data labelling or image generation algorithms producing different results by gender.
Gender bias in data collection, sampling, and grouping	Face recognition programs producing different recognition rates for white males and black females. Male-biased medical data.

Modelling such as algorithm generation and learning	
Algorithm design not considering negative gender bias and its alleviation	Gender gap in exposure to career development advertisements in 'STEM'(Science, Technology, Engineering, and Math). Model training based on computer perspectives strengthening traditional gender stereotypes.
Machine learning based on gender-biased data	Gender gap in exposure to advertisements for high paying jobs (male-biased). Unconditionally deducting points in connection with women-related keywords in the process of AI-based employment.
Difficulties in ensuring the transparency and explainability of algorithms	Giving lower credit to wives than husbands, who are co-owners of properties. When searching via keywords regarding women of color, suggestive content is excessively exposed (at the top of the list, etc.).

○ The analysis of gender bias cases shows that bias occurs in the AI technology configuration as a whole. In other words, it is necessary to consider the issue of gender bias in AI by phase and to develop useful tools for mitigating the bias. Also, fundamental factors that make it difficult to develop and continuously maintain substantial methods to improve situations are deemed to be historical and structural gender bias accumulated and its impact and the failure of tech firms and society as a whole to recognize such bias as a key issue.

3. In-Depth Interviews for Reducing Gender Bias in AI

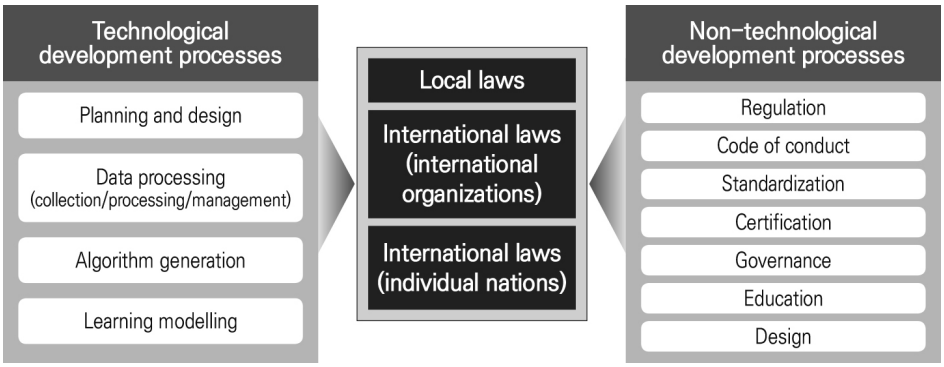
○ In-depth interviews with ten (six males and four females) experts with experience in AI development were conducted for this study. They consist of a questionnaire survey (phase 1), in-depth interviews regarding socio-cultural environments (phase 2), and a survey on competencies for and recognition of bias mitigation technologies (phase 3).

○ The main cause of gender bias in AI was found to be lack of attention to data collection, sorting, and processing. This means that final decision makers (administrators) for AI configuration are fundamentally responsible for the issue. However, because it is difficult to apply bias mitigation technologies and elements due to limited development time and costs, it is necessary to develop 1) a gender bias checklist in AI ethics, 2) gender bias mitigation guidelines for engineers in the phase of AI technology development, and 3) gender bias mitigation guidelines for developers by AI technology in order to reduce gender bias in the realm of technology.

○ It is recognized that ‘the significantly low percentage of women among AI researchers and developers in their 30s’ and ‘the accumulation of experience in gender discrimination,’ rather than gender itself, can impact development processes. In particular, in the field of advanced science and technology, gender-discriminatory working environments are found to be one of the main causes of women’s career interruption. Therefore, it is necessary to identify elements for desirable working environments including gender-balanced human resources (HR) cultivation and thereby to improve relevant systems.

Ⅳ. Local and Global Reactions to Gender Bias in AI

○ AI-related laws and policies in Korea and other nations (including international organizations) are analyzed in the context of technological and non-technological development processes[3)].



[Figure 3] Reactions to Gender Bias in AI by Technological and Non-Technological Development Process

○ Local AI laws and policies are designed to prevent discrimination and bias from occurring in each phase of AI technology development. However, most laws do not have provisions explicitly indicating gender issues.

〈Table 2〉 Local Reactions

Laws	
Technological development processes	Absence of provisions for preventing gender discrimination or controlling gender bias.
Non-technological development processes	The Framework Act on Intelligent Informatization (law 17344) contains comprehensive clauses regarding inequality and gaps, failing to explicitly identify gender bias and discrimination.
Policies	
Technological development processes	AI R&D Strategies for I-Korea 4.0 are established without policies regarding specific development processes such as data processing, algorithm generation, and learning modeling, failing to sufficiently consider gender issues.
Non-technological development processes	Lack of gender-related provisions

3) Technological and non-technological development processes are connected to ‘AI planning, development, and generation’ and ‘AI related human rights and ethics,’ respectively.

Bills sponsored by legislators	
Technological development processes	'The Bill on Algorithms and AI' contains explicit provisions concerning the prevention of gender discrimination.
Non-technological development processes	'The Bill on AI R&D, industry promotion, ethical responsibilities, etc.' contains explicit provisions regarding the protection of human rights and dignity, lacking gender-equal perspectives.
Policy guidelines	
Technological development processes	Ministry of Science and ICT: Ethical Guidelines for an Intelligent Information Society National Human Rights Commission: Human Rights Guidelines for AI Development and Utilization Korean Women Link: AI Guidelines Sponsored by Feminists
Non-technological development processes	National Human Rights Commission: Human Rights Guidelines for AI Development and Utilization Korean Women Link: AI Guidelines Sponsored by Feminists

○ International organizations such as the OECD, the EU, and UNESCO prepared recommendations regarding AI ethics, highlighting that each nation should ensure reliability and consider AI ethics in the process of developing and utilizing AI technologies. As the legislation cases of major nations in the world illustrate, plans to assess and disclose the ripple effects of AI on racial and gender discrimination in a transparent manner should include gender bias.

○ Therefore, the Framework Act on Intelligent Informatization, a fundamental system for Korea's AI technologies, needs to contain provisions concerning gender bias mitigation. Also, learning lessons from UNESCO's Recommendation on the Ethics of Artificial Intelligence, gender bias mitigation content should be reflected in installing and operating a dedicated organization that evaluates and monitors the ethical impacts of AI.

Ⅴ. Policy Suggestions for Mitigating Gender Bias in AI

○ With the rapid development of AI technologies, keen attention has been paid to relevant socio-economic realms. AI's decision-making is widely believed to be neutral and free from prejudice or bias because it is a machine. However, given the current level of technology, AI should be considered as a statistical analysis tool for presenting the best possible solutions under given constraints, rather than as an independent decision-maker.

- Therefore, coming up with solutions to gender bias in AI requires perspectives and endeavors to improve situations, recognizing cyclical links between humans and environments. At this time, people's biased recognition of gender should also be considered, as well as social environments impacting such attitudes.

○ Two types of reactions should be considered to alleviate gender bias in AI.

- First, in terms of technology and engineering, measures to reduce gender bias should be devised in each phase of technology configuration and utilization. As mentioned earlier, gender bias in AI may happen in each phase of AI technology configuration. As a result, in each phase of AI technology configuration including AI planning/design, data processing (collection/processing/management), algorithm generation, and learning (modeling), the issue of gender bias in AI should be taken into account, developing technical guidelines for mitigating such bias.
- Second, gender bias in AI fundamentally comes from gender stereotypes and discrimination in society as a whole. Therefore,

recognizing gender bias in AI as a key issue and learning lessons from AI-related legal systems and policies discussed in international organizations, social environments and relevant policies should be established and developed, respectively, to enhance legal and institutional systems and mitigate gender bias in society.

〈Table 3〉 Policies for Mitigating Gender Bias in AI

1. Improvement plans for technological development processes	**AI technology and system establishment**
	(1) Developing and distributing checklists for assessing ethical and gender bias in AI. (2) Developing and distributing gender bias mitigation guidelines for engineers at the phase of AI technology development. (3) Developing and distributing gender bias mitigation guidelines for developers by AI technology.
2. Improvement plans for non-technological development processes	**Legal and institutional system improvement**
	(1) Including a provision explicitly specifying that gender issues should be considered in the Framework Act on Intelligent Informatization (Article 3: Basic Principles) (2) Including a provision explicitly specifying the gender impact assessment in the Framework Act on Intelligent Informatization (Article 56). (3) Including gender bias prevention in the AI Ethical Impact Assessment. (4) Establishing and operating an AI ethics review organization from gender perspectives. (5) Operating an AI gender bias monitoring group at the government level. (6) Making laws concerning criteria for verifying gender bias in data and algorithms.
	Establishment of social environments
	(1) Providing education on ethical and gender bias in AI. (2) Cultivating and supporting talented females to develop their careers in AI and basic science.

References

AI in Naver Encyclopedia (https://terms.naver.com/entry.naver?docId=1136027&cid=40942&categoryId=32845, Accessed: 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: March 17, 2022).

Yang Jong-mo (2017). Impact of Bias and Opaqueness in AI Algorithms on Legal Decision-Making and Relevant Regulation Plans. The Korean Legal News, June 2017.

Moon Mee-kyung et al. (2022). Gender Bias in the Use of Artificial Intelligence and Deep Learning and Suggestions for Improvement. Korean Women's Development Institute Research Report-11, Republic of Korea.

Heo Eu-sun (2018). Preliminary Consideration for Discussing AI-Induced Discrimination and Its Responsibility – Focus on Learning Bias in Algorithms and Human Actors. Korean Feminist Philosophy 29.

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

KWDI

KWDI
225 Jinheung-ro, Eunpyeong-gu (1-363, Bulgwang-dong) Seoul, 03367, Republic of Korea
TEL 02.3156.7000 FAX 02.3156.7007
http://www.kwdi.re.kr