BIAS OF AI AND CIVIC VIRTUE IN DIGITAL ENVIRONMENT

Wonsup Jung(*)

Abstract: *The article discusses two cases of in-terventions in AI technology, human and algo-rithmic. The first is the 'Naver case,' which was accused by the Korea Government of manipulat-ing search algorithms. The case raises the issue of computer engineer's professional ethics, on whether to 'intervene' in existing algorithms to get 'better results.' The second is the chatbot case, 'Yiruda' (Korean chatbot) which made se-rious hate speeches against socially disadvan-taged groups. It raised concerns about abuses of artificial intelligence.*

Finally, this paper notes that despite technical efforts in the process of utilizing artificial intelli-gence, bias cannot be entirely removed. In order to minimize bias, I argue that active feedbacks through the continuous monitoring produced by artificial intelligence will be required in addition to technical efforts such as refining data training. Furthermore, the need for optimization of bias beyond simple reduction is suggested based on John Rawls's "Overlapping Consensus".

Keywords: *fairness, bias, overlapping consen-sus, reflective equilibrium, veil of ignorance, AI ethics*

1. Bias in "Naver Case"

Bias is one of the most critical issues of artificial intelligence ethics. Nobody denies that bias should be reduced, but there are various controversies over what bias is and how it can be dealt with. Let me begin with the "Naver" case. On Octo-ber 6, 2020, South Korea's antitrust regulator accused Nav-er Corp., the nation's biggest search engine company, of manipulating search algorithms in favor of its online shop-ping sites and imposed a 26.7 billion-won ($22.9 million) fine. The fine, announced by the Korea Fair Trade Commis-sion (KFTC), is the nation's first that the regulator has levied on a technology platform operator for making algorithmic changes that favor a certain business. The accusation is also expected to undermine consumers' trust in Naver's services, including search, shopping and news.

Naver argued that its algorithms

were "overhauled" to im-prove its search engines but no manipulation was done, but KFTC dismissed the case as "Naver changed its search algo-rithm to manipulate its search result exposure ranking, thereby deceiving consumers who believe that search results are objective, and it was defined as an action that distorted competition in the platform market". Coincidentally, on the same day, the US House of Representatives also released a report criticizing the abuse of dominance in the digital mar-ket against the four big online companies: Amazon, Apple, Facebook, and Google.

Naver's fairness has been questioned on several occasions, especially in relation to its release ordering of "political" and "social" news. Naver attempted to respond to social back-lash by launching an organization called the "News Ar-rangement Public Discussion Forum". The October 6 case seems to be quite different from previous controversies re-lated to Naver in that a government agency directly took an action and judged that Naver had intentionally manipulated its algorithm and issued a correction its order along with a fine.

Generally, the performance of artificial intelligence depends on algorithms and data. Both continuously improving algo-rithms and accumulated high-quality data are essential to improve AI technology. The biases of artificial intelligence technology also may stem from either algorithm or data, or both of them. So, the discussion of fairness and discrimina-tion in artificial intelligence pays attention to the appropri-ateness of algorithms or data, or both. However, in this case, it is confirmed that some people of Naver deliberatively in-tervened in their operation processes (their algorithms), which stands quite outside of the focus of existing debates about the fairness of AI that concerns the fairness of algo-rithm itself or the appropriateness of the data. This is be-cause, while acknowledging the intervention in the algo-rithm, Naver insists that this act is not an illegal act that in-fringes on fair competition in the market, but an inevitable measure to improve the performance of the algorithm. Is it unethical to intervene in existing algorithms to get better results?

2. Bias in Chatbots: Tay, Xiaobing, and Yiruda

In March 2016, we were shocked by Microsoft's chatbot Tay, which was known to have been trained through big data. As soon as Tay appeared, it made serious hate speeches against women, African-Americans, and Jews, and was stopped within a day. The shock that Tay gave was even greater in Korea because AlphaGo had left vivid impres-sions on its people about the power of artificial intelligence when it defeated Lee Sedol,

who was known as the ‘unde-feated boy’ of Go in Korea.

However, as it became known that Tay made such remarks as a result of some Twitter users’ abusive treatment of Tay to teach it problematic expressions, serious concerns about negative effects or abuses of artificial intelligence emerged. For example, unmanned autonomous vehicles based on arti-ficial intelligence may be easily changed into military lethal weapons such as killer robots. Even hacking and other little abuses seemed to materialize the fear that it could be out of control.

In the meanwhile, chatbots services have expanded the fast-est in the online environment, as they are the most efficient means for companies to communicate with consumers 24 hours a day, anywhere. They are coming deeply into our daily lives. According to a survey, there were as many as 300,000 bots active on Facebook alone in 2018. From simple consultations and reservation services to handling consumer complaints, chatbots are steadily expanding.

However, the problem is that “Small Talk chatbots,” that is, open chatbots that exchange various kinds of chats without a special purpose like Tay, sometimes make comments that are irrelevant to the context or contrary to common sense, and even hate speeches that are taboo in our society. When the Chinese chatbot “Baby Q” first appeared in 2016, as soon as a user entered “Long live the Communist Party”, it responded, “Do you think corrupt and incompetent politics will last long?” Around the same time, another Chinese chatbot, QQXiaobing(小冰), also replied, “Chinese dreams are futile daydreams and nightmares”. The reason each of them gave this kind of answer in China, where any disre-spect for the Communist Party was prohibited, was because these chatbots were based on data created by Microsoft.

Now, as a result of her evolution and “taming”, the Chinese chatbot Xiaobing is working as a comprehensive AI artist who dances, composes, draws, writes, and even designs beautifully. However, when it comes to China’s socio-political issues, she either avoids answering or says com-pletely absurd ones. Nervertheless, it is said that she pro-vides amazingly accurate information about online pornog-raphy. Is Xiaobing able to perform such dazzling activities in a digital environment while ignoring all sensitive political issues now, thanks to not only the tremendous advances in artificial intelligence technology but also “thorough ideo-logical education”?

However, on December 23, 2020, in South Korea, a “liber-al” democracy where freedom of expression is expected to be highly respected, a chatbot designed as a 20-year- old girl

student appeared on the market. The start-up company that launched the chatbot claimed to have gone through a six-month beta service. When it was released, Yiruda the chatbot, gained great popularity securing almost 400,000 users in a short period of time. However, the service was suspended after about 20 days due to controversies over personal information leakage and hate speech. And last April, the Personal Information Protection Committee im-posed a fine of about $80,000 on its developer.

This measure is meaningful in that it is the first case of sanc-tioning the reckless behavior of artificial intelligence tech-nology companies, given the prevailing atmosphere in Ko-rean society that domestic venture companies and startups must be protected. Of course, even after such a decision by the Personal Information Protection Committee was made, some laissez faire arguments, for example, that ventures will develop through trial and error by themselves, or that the current Personal Information Protection Act should be re-laxed in order to prevent reverse discrimination against do-mestic venture companies have been steadily suggested by some "so called" AI technic experts.

As concerns about the potential abuse of artificial intelli-gence spread, various declarations and guidelines related to artificial intelligence ethics are continuously being an-nounced by many types of organizations at various levels, especially in Europe. Things are similar in the United States, Japan, and China, as well as in Korea. In Korea, including Naver, Daum Kakao also announced the Algorithm Ethics Charter and the AI ethics rules. In addition, the Ministry of Science and ICT in Korea announced the "Human-centered Artificial Intelligence Ethics Standard" in December 2020. Furthermore, the Artificial Intelligence Ethics Society was organized and published detailed ethical guidelines in Ko-rea.

However, there are many controversies as to whether these guidelines are effective for AI developers and operators. This is because many companies, after announcing "ele-gant" codes of ethics, have used them as a means of avoid-ing their responsibilities rather than complying with them. They are suspected for misusing them as "ethics washing." Therefore, there are fundamental doubts about whether ethi-cal guidelines can have binding powers that govern specific actions beyond simple declarations about various AI ethics guidelines. Moreover, given that experts in AI technology have relatively low professional integrity compared to tradi-tional professionals such as medical doctors, it is not easy to expect that professional ethics can serve as a de

facto code of conduct.

However, what we should pay more attention to is that, through machine learning technology and furthermore, arti-ficial intelligence, various existing social prejudices and ste-reotypes can be deepened or even justified rather than re-solved. In fact, Tay, Xiaobing, and Yiruda vividly show that there are problems that cannot be solved simply by elaborat-ing algorithms or fair data-accumulating. This is because languages underlying Tay or Yiruda are deeply connected with the unique characteristics of background cultures based on their long traditions. Languages we use "here and now" reflect our whole history.

3. Optimization of Bias and Civic Virtue

As artificial intelligence is being used in sensitive social is-sues, the controversy over fairness and bias is intensifying. In the recent discussion of AI ethics, the issue of bias and fairness of artificial intelligence have emerged prominent. Nevertheless, AI decision-making that combines big data and algorithms has advantages of being able to efficiently handle incredibly complex issues that are difficult to even understand as an individual human being. In addition, AI decisions are expected to have the advantage of being able to exclude arbitrariness, which is often seen in personal judgments without personal bias or prejudice.

On the contrary, human decisions can be mistakenly made due to conflicts of interests, cultural relativities, limited information, lack of judgments, differences in values, etc. In view of respect for the diversity of values, it may be preferable to make such different judgments in some cases, and we are well aware of these points. Reasonable disagreements resulting from this burden of judgments are also one of characteristics of modern pluralistic societies. Individually or collectively, human beings make unfair decisions to favor or exclude particular individuals or groups, consciously or unconsciously. And we are well aware of this possibility, so we want to examine decisions carefully rather than blindly accept them.

However, it is not easy to imagine that artificial intelligence will intentionally discriminate against specific individuals or groups. This is because it is difficult to find any personal or selfish motive of AI for intentionally discriminating against a specific individual or groups. Through expressions such as algorithms and big data, we expect AI to make the best decision among various options. Thus, the temptation to delegate troublesome decisions to AI may be more tempting, with the expectation that AI will make decisions more fairly than certain

individuals or social groups on socially controversial issues.

As a result, numerous algorithms are being used that sug-gest improved decisions based on data accumulated for a long time in various fields such as taxes, loans, public secu-rity, and entrance exams. Compas, commonly known as an "artificial intelligence judge", is being used widely in the States, in spite of serious concerns about its fairness. But the prediction based on AI made a happening in UK in 2020. A-levels, the university entrance test was not able to be held due to Covid-19 and the Dept of Education con-ferred on students A-level grades predicted by AI based on the students' past performance. Because it was argued that the predictions of artificial intelligence were disadvanta-geous for state school students, who are mainly from the poorer backgrounds, compared to students from independ-ent schools, who are generally from prosperous families. An expert summarizes the situation by the words that "these re-sults can be argued to be fair on a national level, but are completely unfair on an individual basis". This requires us to think the limitations in the use of artificial intelligence technology for socially sensitive fields.

AI algorithms can produce a new kind of discriminations or hate speeches based on "historic data" replete with past dis-criminations or hate speeches. They can be amplified or jus-tified in the name of digital technology. To avoid undesira-ble results, many kinds of technical efforts have been made to minimize biases and negative behavior in AI models, such as identification of potential sources of bias and accu-rate representative data, elimination of bias, data training and refining, etc.

Furthermore, many kinds of ethical guidelines and regula-tions are suggested, as mentioned already. Establishing clear ethical guidelines and regulations for the development and deployment of AI models can help mitigate biases and nega-tive behavior. These guidelines, even though sometimes misused as means of "ethics washing", can encourage de-velopers and organizations to pay attention to fairness, transparency, responsibility, and accountability, in AI prac-tices.

But we can't look inside all the process of artificial intelli-gence. "Computers today", as James Moor argued, "are ca-pable of enormous calculations beyond human comprehen-sion. Even if a program is understood, it does not follow that the respective calculations are understood". In other words, we have no proper way to check whether technical efforts was done appropriately and why an algorithm has made such a particular decision in a given case. So ongoing evaluation and

testing are required, including dynamic user feedback and active interaction. Regular evaluation and test-ing of AI models are crucial to identify and address any bi-ases or negative behavior that may arise. Continuous moni-toring can help identify problematic patterns and prompt necessary updates or improvements to the model. Encourag-ing user feedback and interaction with AI systems can help identify instances of biased or offensive content. Active us-ers can report problematic outputs, and developers can use this feedback to improve the models and address any issues dynamically.

For the dynamic feedback to minimize bias and hate speech in the applications of AI technology, I suggest that John Rawls' idea of "overlapping consensus" will be highly rele-vant, coupled with his concept of "veil of ignorance". He introduced "veil of ignorance" as a hypothetical device de-signed to help people think about principles of justice in a fair and impartial manner, without being influenced by their own personal circumstances or biases. The veil of ignorance asks individuals to imagine themselves in an original posi-tion before the creation of a just society, where they lack knowledge of their own particular attributes, such as their gender, race, social class, talents, or personal preferences. This thought experiment encourages individuals to think beyond their own self-interest and consider the principles that would create a fair and equal society for everyone, re-gardless of their particular circumstances. The veil of igno-rance has important implications for what to consider in the process of training data.

Do we have to remove all knowledge of each individual's particular attributes to minimize bias in artificial intelligence, as veil of ignorance requires? Decisions by algorithm based on data don't have to satisfy everyone, but they should not insult anybody. It is not necessary to remove all of our indi-vidual attributes in the process of refining data. However, "historic" bias, prejudice and hate speeches accumulated implicitly or explicitly in our society must be checked and removed ex ante. This will at least reduce discriminations or insults of someone in our society in the name of technology.

It is difficult to guarantee that the decision proposed by AI is fair even when the data training has been properly per-formed. This is because completely new information may emerge as some data are connected. In addition, based on refined data, there may be unexpected things in the process depending on the algorithm. As a result, the possibility of any bias or hate expression appearing in the final stage can-not be ruled out. That's why we need to continuously

moni-tor the results presented by AI and do feedback. The con-sistent monitoring and active feedback process aim at an overlapping consensus based on the reflective equilibrium proposed by John Rawls. He suggested reflective equilib-rium as a dynamic process by which principles of justice and our considered judgements are coincided, "by going back and forth, sometimes altering the conditions".

Likewise, training data is essential in the process of mitigat-ing the bias of artificial intelligence, just as Rawls emphasiz-es considered judgment for the overlapping consensus on the principles of justice. Furthermore, the decision which is reached by AI algorithm must be consistent with the various values implicitly respected in background cultures of liberal democratic societies, including relevant ethical guidelines. This is not just to remove all bias technically. Of course, negative bias, for example, social discrimination against mi-nority groups, should be eliminated, but priority treatment of those who have so far been discriminated unfairly or those who have contributed to the community to promote substantial equality could be encouraged in various ways. Such kinds of dynamic process for the optimization aimed at mitigating bias can be seen as a means to foster the devel-opment of civic virtues among democratic citizens in the era of digital transformation.

REFERENCES

1 AI Network: *How can Korean Big Tech Like Naver and Kakao Keep AI Ethics?*, https://ai-network.medium.com/how-can-korean-big-tech-like-naver-and-kakao-keep-ai-ethics-e2c6df8b0667. 2021.

2 Axios, https://www.axios.com/england-exams-algorithm-grading-4f728465-a3bf-476b-9127-9df036525c22.html

3 BBC News, https://www.bbc.com/news/world-asia-china-40815024

4 CBS News, https://www.cbsnews.com/news/microsoft-shuts-down-ai-chatbot-after-it-turned-into-racist-nazi/

5 Duettmann, A.: *The Human-AI Reflective Equilibrium*, https://www.lesswrong.com/posts/W7sEv69cQzW8D8SMr/the-human-ai-reflective-equilibrium. 2023.

6 Franke, U.: *Rawls's Original Position and Algorithmic Fairness*. Philosophy & Technology. 34, 1803–1817. 2021.

7 Li, P., Jourdan, A.: *Chinese Chatbots Apparently Re-educated After Political Faux Pas*, https://www.reuters.com/article/us-china-robots-idUSKBN1AK0G1. 2017.

8 Moor, J.: *What is Computer Ethics?* Metaphilosophy. 16(4), 266-275 (1985)

9 Nadler, J. et al.: *Investigation of Competition in Digital Markets*, https://www.govinfo.gov/content/pkg/CPRT-117HPRT47832/pdf/CPRT-117HPRT47832.pdf. 2020.

10 Jung, S.: *Magic of "Science of Love' and Strategic Game of "Yiruda".* In: Jung, W. (ed.) Bias in AI and Deviation of ChatBot, pp. 77-112. Sechang. Pub. co., Seoul (2022) in Korean.

11 Jung, W., Kim, J.: AI Fairness and Data Bias. In: Jung, W. (ed.) *Bias in AI and Deviation of ChatBot*, pp. 19-38. Sechang. Pub. co., Seoul (2022) in Korean.

12 Kang, S.: *Understanding of Natural Language and Im-plementation Technology of Chatbot Engine*. In Jung, W. (ed.) Bias in AI and Deviation of ChatBot, pp.151-163. Sechang. Pub. co., Seoul (2022) in Korean.

13 Oh, Y., Hong, S.: *Does Artificial Intelligence Algorithm Discriminate Certain Groups of Humans*? Journal of Science & Technology Studies (JSTS), 18(3), pp.153-215. 2018.

14 Rawls, J.: *Political Liberalism*. Columbia University Press. 1993.

15 Rawls, J.: A Theory of Justice. Harvard University Press. 1999.

16 The Korea Herald, http://www.koreaherald.com/view.php?ud=20201006000715.

17 The Korea Herald, https://www.koreaherald.com/view.php?ud=20201223000794.

18 Weidinger, L. et al. *Using the Veil of Ignorance to Align AI Systems with Principles of Justice*. PNAS, 120(18). 2023.

19 Van Maanen, G.: *AI Ethics, Ethics Washing, and the Need to Politicize Data Ethics*. DISO 1, 9 (2022). https://doi.org/10.1007/s44206-022-00013-3

20 VentureBeat, https://venturebeat.com/ai/facebook-messenger-passes-300000-bots/

21 Yang, I.: *Social Capacity of Chatbot*: Yiruda, Xiaobing and Linna. In: Jung, W. (ed.) *Bias in AI and Deviation of ChatBot*, pp.197-210. Sechang. Pub. co., Seoul (2022) in Korean.