ICRES 2022: 7th International Conference on Robot Ethics
and Standards, Seoul, South Korea, 18-19 July 2022.
https://doi.org/10.13180/icres.2022.18-19.07.012

BIAS AND FAIRNESS IN AI

WONSUP JUNG

School of Liberal Studies, Kyungnam University,
Changwon, Kyungnam 51767, South Korea
E-mail: wonsupj@kyungnam.ac.kr
www.kyungnam.ac.kr

This paper is focused on artificial intelligence bias. As various artificial intelligence algorithms are used in public domains such as taxation, policing, and judicial law, the debate over bias is intensifying at home and abroad despite their effectiveness. Chapter 1 introduces the main issues related to AI bias. Chapter 2 examines the reasons and types of artificial intelligence bias in light of the characteristics of the algorithm itself. In Chapter 3, I will examine some of bias issues of predictive policing algorithms that have been introduced in the United States and are producing controversy.

Concluding my presentation, I suggest several questions for our session as follows: does artificial intelligence make better judgments than humans? Are artificial intelligence algorithms the best and fairest solution to our public affairs?

1. Characteristics of Artificial Intelligence Bias

On October 6, 2020, South Korea's antitrust regulator accused Naver Corp., the nation's biggest search engine, of manipulating search algorithms in favor of the company's online shopping site and imposed a 26.7 billion-won ($22.9 million) fine. The fine, announced by the Korea Fair Trade Commission (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.[a] On the same day, the US House of Representatives also released a report accusing the abuse of dominance in the digital market against four major online companies, including Amazon, Apple, Facebook, and Google.[b]

Naver, which has virtually dominated the digital market in Korea, has already been a subject of constant debate over the bias of news distribution, so this situation has not little implications for the fairness of digital giants.

The solution direction seems clear to some extent in that it is a basic rule violation that the artificial adjustment of the company insiders caused inequality in the market. In this case, as it is a typical deviation or tyranny of large corporations, it is a path that can be solved in many traditional ways, such as punishing offenders, emphasizing professional ethics, or restructuring various laws and systems that ensure fair competition in the market while protecting whistle blowers.

In March 2016, we were shocked by Microsoft's chatbot Tay in a completely different way. Tay, which is known to have learned with big data, has poured out all sorts of hate speeches about the socially underprivileged, such as women, black people, and Jews, within a day after launch. The shock from Tay was even greater because of the tremendous power that Google Deep Mind's AlphaGo showed in the game against Lee Se-dol.

[a] The Korea Herald(2020, Oct. 06), “Naver faces 26.7b-won fine, accused of manipulating algorithms.” http://www.koreaherald.com/view.php?ud=20201006000715&ACE_SEARCH=1.
[b] Jerrold Nadler, et al. Investigation of Competition in Digital Markets (Majority Staff Report and Recommendations, Subcommittee on Antitrust, Commercial and Administrative Law of the Committee on the Judiciary).

However, two issues emerged as it became known that Tay's remarks were the result of some Twitter users actively learning Tay in a specific direction.

The first issue is the possibility that artificial intelligence technology will be systematically abused after being hacked by a specific group. The discussion on ethics and governance regarding Artificial Intelligence is exploding, and various guidelines have been released by EU Council, US government, Asian Countries, various NGOs, and Big Hi-tech Companies.[c]

The second issue is that existing social prejudices or stereotypes may be represented and/or justified or deepened through machine learning technology and artificial intelligence. Generally, whenever a new technology was introduced, we have sincere hopes that the technology would solve existing troublesome problems, as well as concerns that if the technology was extended to other areas, it would derive new problems. Alphgo is fundamentally based on mathematics. In contrast, our language proceeds in a specific socio-cultural context called “here and now”.

The Tay case clearly showed that the “bias” that can emerge when artificial intelligence technology is expanded far exceeds the level of concern for scientists and engineers. This is because the language Tay learned is deeply related to the biases of our society itself.

Moreover, human decisions can be made different from one person to another due to conflicts of interest, cultural relativity, limitations of information, lack of judgment, differences in values, etc., or may differ from person to person on the same issue, and in some cases, there may be errors. In terms of the diversity of values, making such different judgments may be desirable in some cases, and we are well aware of these points.

However, it is not easy to think that technology, especially artificial intelligence technology, will have a conflict of interest, cultural relativity, or differences in values. Thus, even on socially controversial issues, it can be more tempting to delegate troublesome decisions to AI, expecting that decisions made by artificial intelligence technology will be fairer than those of natural persons.

As a result, a number of algorithms have already been used to suggest improved decisions based on data accumulated in various fields such as taxes, loans, policing, and entrance examinations for a long time. Last summer, the British government, which was unable to take the high school graduation test due to the corona crisis, tried to give grades using artificial intelligence technology.[d] This was a happening as it was argued that public school students, mainly the low-income class, were at a disadvantage compared to the private school students, who were children of the wealthy class, as predicted by artificial intelligence. The remarks of Helena Webb, a senior researcher at the Department of Computer Engineering at Oxford, that these results "can be argued to be fair at the national level, but have completely lost fairness on an individual basis" clearly raises another important issue regarding the use of artificial intelligence technology.[e]

In this article, I would like to examine the bias that has begun to emerge as artificial intelligence is used in social issues such as predictive policing.

2. Types of Algorithmic Bias

As is well known, after the 2000s, methods such as deep learning, which dramatically improved the performance of existing neural networks and machine learning, were introduced into the

[c] In 2018, more than 50 celebrities from around the world who are leading AI technology announced a declaration refusing to collaborate with KAIST, raising concerns that KAIST's AI research could lead to research on killer robots. http://news.kmib.co.kr/article/view.asp?arcid=0923929416.

[d] https://www.axios.com/england-exams-algorithm-grading-4f728465-a3bf-476b-9127-9df036525c22.html.

[e] http://www.hani.co.kr/arti/international/europe/959055.html.

field of artificial intelligence and combined with big data. The efficiency and accuracy of the system have increased remarkably, and the application area has expanded.

However, Google Photo' classified black faces as 'gorillas' raised questions about the accuracy of artificial intelligence algorithms, and various social debates including the possibility of racial discrimination. At the same time, the question was raised about the compass algorithm, which is called an artificial intelligence judge. The debate is still seriously ongoing as the claims that the Compass algorithm, which has been widely used in judgments such as sentence, parole, and bail in US courts and prisons, has emerged against blacks.[f]

However, the problem is that these biased results can come from the following reasons without intentional intervention or artificial adjustment.

First, biased results can emerge because of the numerous correlations inherent in big data. In other words, when my spending patterns, learning patterns, Internet search patterns, hospital visit patterns, and phone usage patterns are combined, there can be many correlations.

When machine learning algorithms that learn themselves from patterns are combined, information about me is newly constructed and predictions about my behavior become possible. As a result, decisions made by algorithms combined with big data can deliberately or unintentionally produce discriminatory results.

First of all, deliberate discrimination can occur in algorithmic decision making, even if sensitive categories such as gender, race, and rich and poor are not directly included.

Second, biased results can emerge in the four stages of data mining. Data mining consists of four steps:

First, target variable definition
Second, data labeling and gathering,
third, feature selection
and fourth, decision making based on modelling

In this way, not only when discrimination is intended, but even though there is no explicit intention of discrimination, it is sufficient that bias will be involved in each step of data mining.

If so, is it possible to improve the problem of algorithmic unfairness by minimizing the potential for bias, discrimination, and unfairness at each of these steps?

At first glance, it is a valid reasoning, but several researchers point out that attempts to ensure algorithmic fairness may raise new problems. This is because efforts to solve the problem of algorithm fairness may conflict with efforts to estimate or predict unknown information as accurately as possible, which is the original purpose of big data algorithms. For example, if more detailed data is collected to increase accuracy, the combination of detailed data can eventually lead to a specific layer.[g] The problem of algorithm fairness has a trade-off with the problem of algorithm accuracy.

3. Bias in Predictive Policing

According to a report published by RAND Corporation, there are four types of predictive policing:
predicting a crime,
predicting a criminal,

[f] Angwin, Larson, Mattu, and Kirchner, 2016.
[g] Selbst, 2018: 137; cf. Williams et al., 2018.

predicting the identity of a criminal (from past crimes),
and predicting a victim.[h]

These crime prediction policing usually consists of four steps:

First, data collection step,
second, collected data analysis and prediction step,
third, crime response strategy planning step,
and fourth, criminal response step

In addition, this process continuously “feeds back” and improves the accuracy of predictions, thereby increasing the efficiency of public safety. What is important in predicting crime is where and when the crime is expected. David Weisburd argued that the paradigm of crime should shift from a people-centric paradigm to a place-centric paradigm. A company called PredPol was founded in 2012 to commercialize this technology. The company's crime prediction algorithm, Fred Paul, has garnered great media attention since its launch.[i]

Even if the prediction is not 100% accurate, predictive algorithms have been evaluated to have several advantages. First of all, the algorithm made it possible to focus the patrol power of the police in a specific area or situation. And it provided data to analyze in which circumstances crime is more likely to occur. It allowed them to plan measures to prevent crime more effectively, and could even help police use specific strategies.[j]

However, problems with the Fred Paul algorithm soon emerged.[k] Above all, a problem could be raised about the poor quality of data used by the algorithm. Most seriously pointed out was the criticism that these programs violate the right to citizenship, particularly the privacy set out in the Fourth Amendment. Indeed, within a few years, this criticism poured out from academia, legal circles, and human rights groups.

O'Neill criticized that it was wrong to include misdemeanor crimes in the data when creating the Fred Paul algorithm.[l] As data on misdemeanor crimes were included, areas where many blacks lived were classified as misdemeanor areas from the beginning. Police patrol these areas, and then they stop and check people who appear suspicious. Among them, people with drugs or suspicious objects were arrested. These arrests are fed back into this algorithm and function as data, increasing the success rate of the Fred Paul algorithm. This in turn leads the police to intensive patrol areas where black people live.[m]

Selbst (2018) pointed out that Fredpol's prediction was problematic in four ways.

First, the problem of bias arises in the process of computerizing crime. Crimes come in a variety of ways and there are many types, but prediction programs focus on the crimes that computers can easily recognize.

Second, bias can occur in the process of 'training' big data. Algorithms such as Fred Paul used a racially biased reference value, and it is highly likely that the area where many blacks or

[h] Perry et al., 2013.

[i] The Economist (2013. 7. 20), An Internet Archive page of Kent Police (2013. 12. 3) ‘‘Predictive Policing day of action targets burglars,’’ archived on May 2nd, 2014.

[j] Schlehahn et al,, 2015.

[k] Perry et al., 2013: 12-13.

[l] Cathy O’Neil. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy,* 2016.

[m] O’neil, 2017.

Hispanics lived from the beginning were identified as criminal areas and concentrated patrol power there.[n]

Third, bias arises in the process of selecting specific features of data.

The final bias is the imbalance between 'accuracy' and 'fairness'. Predictions can be more accurate, but fairness is more compromised. Legal scholars believe that big data and predictive algorithms are leading to further weakening of the 4th Amendment to the U.S. Constitution, which was ineffective against discrimination such as racism.[o] So Selbst proposed a way to limit the use of police. He argued that police using crime prediction algorithms should prepare and submit "Algorithm Impact Statements" (AIS). Just as environmental impact assessment has changed the past perception that pollution and health violations are inevitable for industrial development, algorithmic impact assessment can also change the current perception that minor human rights violations are inevitable for the development of the data industry. His proposal to write an algorithmic impact assessment is a proposal made under the assumption that the algorithm will continue to be used. This assumption is based on another assumption that artificial intelligence algorithms will make better judgments than human judgments.

But does artificial intelligence make better judgments than humans? Are artificial intelligence algorithms better and fairer solution to our public affairs? Or is it cheaper and seemingly more efficient solution?[p]

References

1. M. Barranco and J. R. Buchler, *Phys. Rev.* **Cf22**, 1729 (1980).
2. H. Müller and B. D. Serot, *Phys. Rev.* **C52**, 2072 (1995).
3. V. Baran, M. Colonna, M. Di Toro and A. B. Larionov, *Nucl. Phys.* **A632**, 287 (1998).
4. V. Baran, M. Colonna, M. Di Toro and V. Greco, *Phys. Rev. Lett.* **86**, 4492 (2001).
1. Angwin, J. and Larwon, J., ''Bias in Criminal Risk Scores is Mathematically Inevitable, Researchers Say'', *ProPublica*, (2016). [https://www.propublica.org/article/bias-in-criminal-risk-scores-is-mat hematically-inevitable-researchers-say]
2. Burrell, J., ''How the machine 'thinks': Understanding opacity in machine learning algorithms.'', *Big Data & Society*, **Vol. 3, No. 1**, (2016).
3. Ferguson, A. G., ''Big Data and Predictive Reasonable Suspicion'', *University of Pennsylvania Law Review*, **Vol. 163, No. 2**, pp. 327-410, (2015).
4. Ferguson, A. G., *The Rise of Big Data Policing: Surveillance, Race, and the Future of Law Enforcement*, NYU Press, (2017).
5. Greene, D., Hoffmann, A. L., and Stark, L., ''Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning'', *The 52nd Annual Hawaii International Conference on System Sciences (HICSS)*, Maui, HI, (2019).
6. Hand, D. J., ''Classifier Technology and the Illusion of Progress'', *Statistical science*, **Vol. 21, No.1**, pp. 1-14, (2006).
7. Liu, H. and Motoda., H., *Feature selection for knowledge discovery and data mining*, Springer Science & Business Media, (2012).
8. Michael, M., & Lupton, D., ''Toward a manifesto for the 'public understanding of big data'.'', *Public Understanding of Science*, **Vol. 25, No. 1**, pp. 104-116, (2016).

[n] Selbst, 2018: 133-135.
[o] Perry et al., 2013: 13.
[p] Yoehan Oh and Sung-ook Hong (2018), "Does Artificial Intelligence Algorithm Discriminate?" *Journal of Science & Technology Srudies* 18(3).

9. Jerrold Nadler, et al., Investigation of Competition in Digital Markets (Majority Staff Report and Recommendations, Subcommittee on Antitrust, Commercial and Administrative Law of the Committee on the Judiciary).
10. Noble, S. U., *Algorithms of Oppression: How search engines reinforce racism*, New York: NYU Press, (2018).
11. Perry, W. L., McInnis, B., Price, C. C., Smith, S. C., and Hollywood, J. S., *Predictive policing: The Role of Crime Forecasting in Law Enforcement Operations, Rand Corporation*, (2013).
12. PredPol (n.a.), ‘‘Company | Management Team | About Us | PredPol Predictive Policing’’ http://www.predpol.com/about/company/.
13. Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C., ‘‘Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms’’, *In Data and Discrimination: Converting Critical Concerns into Productive: A preconference at the 64th Annual Meeting of the International Communication Association. Seattle*, WA, (2014).
14. Schlehahn, E., Aichroth, P., Mann, S., Schreiner, R., Lang, U., Shepherd, I. D. H., and Wong, B. L. W., ‘‘Benefits and Pitfalls of Predictive Policing’’, *Proceedings of 2015 European Intelligence and Security Informatics Conference (EISIC 2015)*, (2015). [https://ieeexplore. ieee.org/document/7379738].
15. Selbst, A. D., ‘‘Disparate Impact in Big Data Policing’’, *Georgia Law Review*, **Vol. 52**, pp. 109-195, (2018).
16. Toole, J. L., Eagle, N., and Plotkin, J. B., ‘‘Spatiotemporal correlations in Criminal Oense Records.’’, *ACM Transactions on Intelligent Systems and Technology*, **Vol. 2, No. 4**, pp. 1-38, (2011).
17. Weisburd, D., ‘‘Place-based policing’’, *Ideas in American policing*, **Vol. 9**, pp. 1-15, (2008).
18. Williams, B. A., Brooks, C. F., and Shmargad, Y., ‘‘How Algorithms Discriminate Based on Data They Lack’’, *Journal of Information Policy*, **Vol. 8, No. 1**, pp. 78-115, (2018).
19. Yoehan Oh and Sung-ook Hong, “Does Artificial Intelligence Algorithm Discriminate?” *Journal of Science & Technology Srudies*, **18(3)**, (2018).
20. ACM U.S. Public Policy Council and ACM Europe Policy Committee, (2017).
21. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, (2017).
22. The Economist, “Don’t even think about it - Predictive policing”, (2013. 7. 20).
23. Los Angeles Times, “Stopping Crime Before It Starts”, (2010. 8. 21); Santa Cruz Sentinel, “Modest Gains in First Six Months of Santa Cruz’s Predictive Police Program”, (2012. 2. 26); NPR, “At LAPD, Predicting Crimes Before They Happen”, (2011. 11. 26) http://www.npr.org/2011/11/26/142758000/at-lapd –predictingcrimes-before-they-happen.
24. San Francisco Weekly, “All Tomorrow’s Crime: The Future of Policing Looks a Lot Like Good Branding”, (2013. 10. 30); Santa Cruz Sentinel, “Modest Gains in First Six Months of Santa Cruz’s Predictive Police Program”, (2012.02.26).
25. An Internet Archive page of Kent Police, “Predictive Policing day of action targets burglars”, (2013. 12. 3) archived on May 2nd, 2014. https://www.axios.com/england-exams-algorithm-grading-4f728465-a3bf-476b-9127-9df036525c22.html.
26. Chammah, M. with Hansen, M., “Policing the Future: In the aftermath of Ferguson, St. Louis cops embrace crime-predicting software”, The Verge, (2016. 2. 3) https://www.thever ge.com/2016/2/3/10895804/st-louis-police-hunchlab-predictive-policing-marshall-project.
27. ‘Google Photos, ...’ https://twitter.com/jackyalcine/status/615329515909156865.
28. City of New York (n.a.) ‘‘NYC Open Data’’ https://opendata.cityofnewyork.us/ ; City of New York (n.a.) ‘‘Results matching of Policy Department (NYPD) | NYC Open Data’’

https://data.cityofnewyork.us/browse?Dataset-Information_Agency=Police+Department+%28NYPD%29.