Bias in Artificial Intelligence

Gamze Büşra Kaya | Bilkent University

SECTION I: Introduction

Artificial intelligence systems are reshaping political and social institutions. Data-driven decision-making technologies are increasingly used in the justice system, recruiting, medicine and banking. The application of algorithmic decision-making technologies in various social contexts causes that many ethical problems have emerged. Although some people think that artificial intelligence technologies can be beneficial for the society in terms of justice by avoiding biased decision-making, there is empirical evidence showing the use of AI systems can often replicate historical bias, rather than alleviate them (Zimmermann et al., p.4). In the justice system, for instance, algorithmic decision-making technologies are used to inform decisions about parole, bail, and prison sentencing and they are biased against historically marginalized groups (Angwin et al., 2016). As another example, algorithmic decision-making systems that are used to help to hire decisions are biased based on gender, race, and age, etc.

These biased outputs coming from AI decision systems generally result from biased training data and biases in training data are inherited from humans. Since these biases are inherited from humans to algorithms, they are not just technical issues, and their solutions cannot be just technical. To better understand why we cannot solve the problem with technical solutions alone, we can look at ProPublica's study. This study showed that the algorithm that is used for rating a defendant's risk of future crime is more likely to give a high-risk score for a black defendant who is actually low-risk than a similar white defendant (Angwin et al., 2016). Since this algorithm is fed socio-demographic information, its outcome is biased against historically marginalized groups. At this point, removing the socio-demographic information from the data can be considered as a solution, but removing it from data to decrease the risk of bias makes data less accurate, which makes outcomes less accurate. To solve this problem, secondly, enforcing equal impact between black and white people in recidivism prediction systems can be offered, but if black people have higher re-offending rates than white people, it causes that white people receive more prison sentences although they are less likely to re-offend. Still, there is a trade-off between the accuracy of outcomes and the risk of bias. Optimizing data without considering the social context that the algorithmic decision-making system used in is not a sufficient solution because there is a need for ethical discussion about bias and justice. Without these discussions, technical solutions are insufficient. Also, since each algorithmic decision- making technology is used in a different social context, they may require different ethical discussions and solutions. For this reason, in this paper, I will focus on only algorithmic decision-making systems that are used in recruiting.

In recruiting process, algorithms are used to screen résumés and evaluate personality tests. These algorithms are so efficient that 72% of résumés are weeded out before a person sees them (O'Neil, 2017). The aim of these algorithms is making discrimination between job candidates, so they are biased in the sense of being discriminatory. For instance, the résumé screening must be biased in favor of the candidate's characteristics best suited to the position.

Whether a particular discrimination is unjust and unfair is an ethical and political question (Coeckelbergh, p.163). The main issue is that the algorithms used in the recruiting process can be biased in an unjust way. For instance, the algorithmic decision-making system used by Amazon did not rank candidates for software developer jobs and other technical positions on a gender-neutral basis because Amazon’s system was trained to examine applicants by observing patterns in résumés sent to the company over a 10-year period (Dastin, 2018). Due to the male dominance in the tech industry, the résumés came mostly from men, which results that Amazon’s system assumed that men are more suitable and preferable for hiring than women. As can be seen from this example, the algorithms used for recruitment could develop an unintended morally wrong bias against historically marginalized groups by inheriting human bias. To deal with this morally wrong bias, determining what kind of bias that decision-making algorithms have is important. Because algorithmic decision-making used for recruiting could be considered as a form of generalization on steroids, any explanation of discrimination based on the wrongness of generalization would be particularly relevant to our current concerns (Binns, p.546). At this point, statistical discrimination account can be used to explain to what extent and under what circumstances wrong the bias in decision-making algorithms, and what kind of solutions can be offered for it. Statistical discrimination is used to understand the characteristics of members by using statistical generalizations concerning groups. For instance, a business owner thinks that s/he has evidence that women with children are generally less focused on work than women without children or men, which leads that s/he declines the job application of a woman with children and accepts the job application of a less qualified person whom she thinks might be more focused on her job.

The main purpose of this paper is to identify in which situations and why the biases of algorithmic decision-making used in recruitment are morally wrong and to find a solution to this morally wrong bias. Because algorithmic decision-making systems are produced by applying rigorous statistical methods, their biased decisions have a higher degree of epistemic warrant than humans' biased decisions (Castro, p.409). Some biases of algorithmic decision-making systems based on rigorous statistical analysis are epistemically reliable and morally permissible. For this reason, firstly, I will focus on statistical discrimination. I will argue that although statistical discrimination is morally right in itself, there are some morally wrong statistical discriminatory practices due to diverse contingent reasons such as having no sound empirical basis and having erroneous beliefs. After that, I will explain potential biases of algorithmic decision-making systems used for the recruitment like in Amazon's system include morally wrong statistical discrimination. Amazon's system's usage of gender as a proxy originated from erroneous belief, which makes it biased in a morally wrong way. I will argue that the making decision-making algorithm used in recruitment should be made blind to gender and race. For this, however, some necessary conditions should be met in some cases. If a trait is observable or measurable, closely related to job performance, and using it in the algorithm is morally permissible, it should be used in the algorithm instead of gender and race. However, algorithms may adopt proxy attributes that correlate with the socially-sensitive attributes such as zip codes that can be used for labeling 'African American' (Johnson, p.12-13). For this reason, this solution is partial. So, making algorithms as much as possible blind to attributes that correlate with gender and race is important.

SECTION II: Statistical Discrimination

Discrimination, generally, relies on the fundamental distinction between people or groups of people which the discriminator makes. Some types of discrimination are instruments of some goals. For instance, young people must be of legal age to take certain types of actions such as driving a car or drinking alcohol in most countries. It seems that the discrimination, in this case, is justified since its goal is to discriminate responsible ones from irresponsible ones based on the reliable assumption that people over the legal age are more responsible than people under the legal age.

Statistical discrimination is that the discriminator relies on a statistically valid empirical distinction between people or groups of people (Schauer, p.43-44). It is used to follow a legitimate end and provides instrumental benefits. In statistical discrimination, the discriminator's aim is independent of discrimination. The discriminator aims to get the most beneficial outcome based on the statistically valid empirical distinction rather than being hostile to a group or considering a group as inferior. Although the discriminator consequently discriminated for her aim, the underlying aim is actually legitimate, so the distinction takes the role of proxy for something else. For instance, insurance companies charge young men higher premiums for car insurance based on the fact that young men are involved in more car accidents. Here, the companies are not hostile to young men or do not consider them as inferior. They just make a distinction for the most efficient outcome. To be more explicit, assume a scenario that there is no statistical evidence that people belonging to the X race are involved in more car accidents, but insurance companies still charge the members of the X race higher premiums for car insurance. In this scenario, it is highly probable that the companies are hostile to members of the X race, so they do not have a legitimate reason to charge members of the X race higher premiums. However, in the previous instance, the reason that companies charge young men higher premiums is not hostility, but the relevant statistical evidence. It seems that statistical discrimination is not morally wrong in itself. In fact, it is a tool that everyone uses (has to use somehow) in daily life.

However, some contingent reasons might make the statistical discrimination morally wrong. Without having a sound empirical basis, making statistical discrimination based on groundless correlations is unjustifiable. For instance, people believing astrology thinks that Capricorns are ambitious and a business that wants to hire ambitious people might prefer people with Capricorn zodiac sign, but there is no evidence to prove or refute this belief (Schauer, p.47). Such statistical discrimination is spurious since there is no sound empirical basis. Likewise, it has been observed that women are not recruited to certain types of jobs due to spurious statistical discriminations against women seen many times in history. These statistical discriminations are based on the false belief that being a woman is a statistically reliable predictor of being worse at dealing with those types of jobs. Therefore, although statistical discrimination is not morally wrong in itself, it might be morally wrong due to some contingent reasons such as having no sound empirical basis or erroneous beliefs.

SECTION III: Morally Wrong Statistical Discrimination in Algorithmic Decision-Making Systems

Because algorithmic decision-making systems are produced by applying rigorous statistical methods, their biased decisions have a higher degree of epistemic warrant than humans' biased decisions (Castro, p.409). In other words, algorithmic decision-making systems' some biases based on rigorous statistical analysis are epistemically reliable and morally permissible. However, some biases in algorithmic decision-making systems stem from morally wrong statistical discrimination. Since I have focused on the algorithmic decision-making systems used for the recruitment, I will explain potential biases of algorithmic decision-making systems used for the recruitment like in Amazon's system include morally wrong statistical discrimination in this part.

Some statistically sound indicators are derived from previous spurious ones and the soundness of these indicators originate from non-statistically justified discrimination, which makes that statistical discrimination morally wrong. For instance, the training data of Amazon's system was comprised of the résumés coming mostly from men, which causes that Amazon's system assumed that being a man is a significant qualification for that job. That is, for Amazon's system, gender plausibly could have been a statistically sound indicator and thus proxy of expertise of software developer jobs and other technical positions. However, the male dominance in the tech industry is based on past bias and discrimination which were being ground on some unfair situations and erroneous beliefs. The instances of these unfair situations might be less educational and economic opportunities for women. The belief that women are less capable in the tech industry might be an example of erroneous beliefs. (More reasons for this bias and discrimination can be given and even their truthness can be discussed. However, this discussion is not directly related to the main claim of this paper, so I will not elaborate on this and will accept the aforementioned erroneous belief as one of the reasons for this bias and discrimination.) Briefly, Amazon's system's usage of gender as a proxy originated from erroneous belief, which means that the system includes morally wrong statistical discrimination. So, using the system for the recruitment process is morally wrong.

Since statistical discrimination that is made by algorithmic decision-making systems used for the recruitment is based on our erroneous belief, it cannot be said that it is epistemically reliable and morally permissible. Therefore, biases of the algorithmic decision-making systems used for the recruitment originate from morally wrong statistical discrimination and they must be eliminated so that the usage of the system can be morally acceptable.

SECTION IV: Making Algorithms Blind to Gender and Race

Since the usage of gender as a proxy in algorithmic decision-making systems used for recruitment leads to unjustifiable statistical discrimination, they must be made blind to gender. Although the example I mentioned above does not include race as a proxy, it is possible that algorithmic decision-making systems use race as a proxy of expertise of any kind of job. Since black people were not preferred for hiring for many years, algorithmic decision-making systems might assume that being white is a significant qualification for the jobs and race might be a statistically sound indicator. For this reason, algorithmic decision-making systems must be made blind to both gender and race so that systems can make decisions without having unjustifiable statistical discrimination against women and black people.

The reason that I offer this solution is that it worked for people. In the 1970s and 1980s, orchestras began using blind auditions, which increased the proportion of women in symphony orchestras (Goldin & Rouse, p.735-736). In addition, it has been shown that a similar example exists in algorithms, although not exactly the same. When a woman's résumé that is rejected by algorithms is showed as a male's résumé to algorithms, it was accepted (O'Neil, 2017).

Yet, making algorithms blind to race and gender might not be possible in all situations; namely, some conditions should be appropriate for this. There must be a trait closely related to job performance, and using this trait in the algorithm is morally permissible. In this way, algorithms can decide whether people are suitable for the job by evaluating the degree to which they have this trait. Although currently there is no evidence for this, there might be an imaginary scenario that in some cases gender and race are correlated with a hidden trait that is closely related to the job performance. In those cases, since hidden traits cannot be observed or measured, gender and race would be used as a proxy. It seems problematic, but the unobservability of hidden traits would force us to do this. Otherwise, it would be impossible to hire the most qualified person for the job. So, making algorithms blind to gender and race seems impossible if there would be a hidden trait that is closely related to job performance. Then, it can be said that using a trait that is correlated with gender or race as a proxy instead of gender and race (makes algorithms blind to gender and race) requires that the trait is observable or measurable, closely related to job performance, and using the trait in the algorithm is morally permissible. Therefore, if a trait is observable or measurable, closely related to job performance, and using it in the algorithm is morally permissible, it should be used in algorithm instead of gender and race; namely, algorithms should be made blind to gender and race.

However, algorithms adopt proxy attributes that correlate with the socially-sensitive attributes such as zip codes that can be used for labeling 'African American' (Johnson, p. 12-13). In that case, making algorithms blind to race is useless because algorithms still can detect race according to the job candidate's zip code. So, this solution might not be useful in all cases, namely, it is a partial solution. For this reason, making algorithms as much as possible blind to attributes that correlate with gender and race is important. Yet, it seems impossible to make algorithms blind to all

References

Binns, R. (2017). "Algorithmic Accountability and Public Reason," Philosophy & Technology, 31(4): 543–556. doi:10.1007/s13347-017-0263-5

Castro, C. (2019). "What's Wrong with Machine Bias," Ergo, an Open Access Journal of Philosophy, 6 (20201021). doi:10.3998/ergo.12405314.0006.015

Coeckelbergh, M. (2020). AI ethics. Cambridge, MA: The MIT Press.

Dastin, J. (2018). "Amazon scraps secret AI recruiting tool that showed bias against women," Retrieved October 31, 2020, from https://www.reuters.com/article/us- amazon-com-jobs-automation-insight-idUSKCN1MK08G

Goldin, C., & Rouse, C. (2000). "Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians," American Economic Review, 90(4): 715–741. doi:10.1257/aer.90.4.715

Johnson, G. M. (2020). "Algorithmic bias: On the implicit biases of social technology," Synthese.

Julia Angwin, J. (2016). Machine Bias.

O'Neil, G. (2017). Hiring Algorithms Are Not Neutral. Retrieved October 31, 2020, from https://hbr.org/2016/12/hiring-algorithms-are-not-neutral

Schauer, F. (2018). "Statistical (and Non-Statistical) Discrimination"

Lippert-Rasmussen, K. The Routledge Handbook of The Ethics of Discrimination. Londres: Routledge.

Zimmermann, A., Di Rosa, E., & Kim, H. (2020). Technology Can't Fix Algorithmic Injustice. Retrieved from http://bostonreview.net/science-nature-politics/annette-zimmerman