Erasmus Institute for
Business Economics

If algorithms are so smart, then why do they discriminate?

Gepubliceerd op: 20-05-2021. © 2021, Erasmus Institute for Business Economics

Bas Donkers is professor of marketing research at the Erasmus School of Economics. In this second article of the 'Data-Economy' series, he investigates the problem of discrimination by algorithms and explores innovative solutions.

The public debate on discriminating algorithms is rising. When algorithms turn out to be better at driving a car, still, it is surprising – and extremely unfortunate – that they tend to discriminate when asked to select promising candidates based on resumes (Dastin, 2018) or those who are more likely to break the rules, which is defined by police profiling (Eligon and Williams, 2015). A key insight into understanding why algorithms have been discriminating – and the good news is that they do not need to – is the fact that algorithms are not designed to be "smart", but they have been developed to be extremely good learners. In fact, what we now call artificial intelligence in many cases stems from the field of machine learning, not machine morality or machine smartness. As a result, applying algorithms on data without additional knowledge on how the data was obtained and how it should be used needs to be done with extreme caution.

In the case of discriminating algorithms, this gives me mixed feelings, as there is some good and some bad news. When algorithms that learn to replicate the complex patterns of human decision-making end up discriminating, this means that the human behaviour that resulted in the observed data to train the algorithm was based on discriminatory behaviour of those humans. If we do not tell the algorithm that it should not discriminate, there is nothing that will prevent it from continuing our own discriminatory behaviour. However, if we can tell the algorithm it should not discriminate, it could well be able to outperform humans in that task.

Discrimination, in the literal sense of the word, means "making a distinction". Discrimination in its negative interpretation is the act of making unjustified distinctions between human beings based on the groups, classes, or other categories to which they are perceived to belong, and subsequently, unfairly treating them in a way which is worse than other people are treated, on the basis of their actual or perceived membership in certain groups or social categories. In the current social and legal context, discrimination comes down to "making unlawful distinctions between people or groups" or "not treating equal cases equally". That algorithms are not by definition free of discrimination has also been shown (see the machine bias example).

Machine bias
There's software used across the country to predict future criminals. And it's biased against blacks.

On a spring afternoon in 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid's blue Huffy bicycle and a silver Razor scooter. Borden and a friend grabbed the bike and scooter and tried to ride them down the street in the Fort Lauderdale suburb of Coral Springs. Just as the 18-year-old girls were realizing they were too big for the tiny conveyances, which belonged to a 6-year-old boy, a woman came running after them saying, "That's my kid's stuff". Borden and her friend immediately dropped the bike and scooter and walked away. But it was too late, a neighbour who witnessed the heist had already called the police. Borden and her friend were arrested and charged with burglary and petty theft for the items, which were valued at a total of $80.

Compare their crime with a similar one: the previous summer, 41-year-old Vernon Prater was picked up for shoplifting $86.35 worth of tools from a nearby Home Depot store. Prater was the

more seasoned criminal. He had already been convicted of armed robbery and attempted armed robbery, for which he served five years in prison, in addition to another armed robbery charge. Borden had a record, too, but it was for misdemeanours committed when she was a juvenile.

Yet something odd happened when Borden and Prater were booked into jail: a computer program spat out a score predicting the likelihood of each committing a future crime. Borden, who is black, was rated a high risk. Prater, who is white, was rated a low risk.

By Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner
ProPublica, May 23, 2016
https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

The good news is that we now have tools to educate selection algorithms that they should not discriminate. The best news here is that algorithms are great learners, which means that they could learn this in a few days, while removing all kinds of implicit and explicit discriminatory biases in human behaviour is likely a matter of months, years or maybe even generations. The intuition behind these debiased algorithms is actually quite simple. A deep learning model (a deep neural network to be more specific) is a sequence of layers with each a number of nodes. When information travels through these layers, the raw information contained in the input/predictor variables is distilled and transformed into the final signal that contains the information needed to best predict the outcome variable. The big success of deep learning is driven by the scope and flexibility of the signal extraction process they provide.

When the signals (in more technical terms the node activations in each layer) represent the information that is traveling through the network to result in a prediction or recommendation, then we might be able to adjust that information to become non-discriminatory. For example, we could tell the algorithm that information about gender is not permitted in the final layers of the network. Once that information is removed, the algorithms final outputs will be unable to discriminate based on these variables, as the necessary information is simply not present anymore.

Telling the algorithm that it is not allowed to use information about gender, is very different from simply removing the explicit information about gender from the data. The problem is that such information is also implicitly present in many other input variables that deep learning algorithms might be using. To forbid the algorithm to use such information in the final predictions, we use its own power to achieve this. Specifically, we use a separate, adversary model that aims to predict gender based on the information contained in at least one full layer of the network. When this model is able to predict gender to only the smallest degree, then there is still some gender-relevant information contained in that layer. At that stage, the model is not obeying our rules and a penalty is imposed on the model to inform the learning algorithm that it is not doing its job properly. In fact, to not discriminate at all, a model that uses gender-relevant information in the final prediction will need to be disqualified, just like an athlete using doping.

Based on these ideas, recently developed algorithms, if properly implemented, are able to select or recommend in a non-discriminatory manner. This does, however, require a definition of discrimination that is understandable to a computer. With many, sometimes conflicting, definitions of discrimination (Kleinberg et al., 2016), the first task is to define with mathematical precision what it means for an algorithm to discriminate. Next one needs to be sure that the variable that one aims to predict should be related to the final decision in a non-discriminatory manner, which is not always the case (Obermeyer et al., 2019). Turning to the data, algorithms should also not be asked to predict far outside the range of observations. If an industry has never hired a female CEO, the data will not be very informative about the characteristics of a good female CEO.

Even with the relevant data, the right model structure and a mathematical definition of non-discriminating behaviour, we might not be there yet. Given that the algorithm is forbidden to discriminate, it will have a worse fit to the data, which is of course only a small sacrifice in order to operate at higher moral standards. In some instances, however, this sacrifice might be unnecessarily large. For example, one might want to differentiate on a small, well defined set of variables that are potentially informative about gender, but where it is morally acceptable to make a distinction that improves fit and accuracy. One specific example that comes to mind is the effect of education level on being qualified for a job, with females being over-represented in the recent cohorts of university graduates. It would be important to reward them for their higher level of education, but if all information that is in any sense informative about gender is to be removed, the algorithm will not be able to fully incorporate the benefits of more education as that would create a gender-related difference. Good algorithms should be able to obey a more nuanced set of (moral) rules.

In current research, we are extending the algorithm such that they do not discriminate based on a specified set of "forbidden" characteristics (e.g. gender), but at the same time are allowed to differentiate based on a second set of "admissible" characteristics (e.g. education). When the latter set of admissible variables is correlated with the forbidden characteristics, the algorithm is allowed to differentiate between men and women, but only based on differences in the admissible characteristics. We have the ideas, and the first results are promising. Feel free to connect when you want to get a heads up once we succeed.

An important caveat here is that this entails the algorithm itself and does not circumvent the age-old adage of garbage in garbage out. If data collection was done in a discriminatory manner, additional measures will need to be put in place to avoid the final outcomes of the algorithm to be discriminating. Hence, non-discrimination already starts at the data-collection stage.

Powerful algorithms tend to work well given the information that is provided to them. However, they, by definition, also lack all the knowledge that humans have but that is not made explicit to the algorithm. To illustrate that data itself is not sufficient for an algorithm to work well, and that additional information on the way the world works or the algorithm should work is needed, consider the relationship between being overweight (or not) and drinking regular versus low-calorie sodas. Suppose data is available on this, which is summarized in Table 1. The algorithm can easily learn the association between soda type and being overweight or not. When asked to recommend a drink for an overweight person, the algorithm can compare the two rows in Table 1 and conclude that overweight people are more likely to drink low-calorie sodas. When asked what to drink in case a person wants to lose weight, the algorithm can compare the weight of individuals who drink regular soda (with the majority of them not being overweight) with those who drink low-calorie soda (with the majority being overweight). The algorithm then would falsely recommend a regular soda to lose weight, as this type of drink is most strongly associated with not being overweight. The algorithm lacks knowledge on the way the world functions and that even though regular soda likely makes you overweight, the pattern is the data is largely driven by overweight individuals choosing to reduce their calorie intake and hence preferring the low-calorie soda.

Table 1: Relationship between the choice of soda and whether someone is overweight

	Low calorie soda	Regular soda
Not overweight	25	75
Overweight	80	20

Data that discriminate

It is possible to remove biases from data that is available. However, it might be much more difficult to correct for biases that result from the absence of data. When there is no female CEO, an algorithm might not be able to determine the features that define a good female CEO. While this is rather obvious, absence of data might also result in discrimination when that data is linked to gender or race. An example is Google discriminating blacks by labelling them as gorillas (Guynn, 2015; Barr, 2015). In this case the algorithm did not receive enough information to differentiate those.

In other domains, measurement instruments might have been devised that lead to discrimination. Approximating a person's health with more easily observed medical expenditures led to white individuals receiving higher priorities as they tended to have higher spending (Obermeyer et al., 2019). This was not driven by their health condition, but by them having better access to health care.

Another area where this is of concern, is HRM, where all kinds of tests have been developed to discriminate the successful candidates from the non-successful candidates. If these tests have been developed based on a male test population, then the instruments might be very ineffective at identifying the successful women as the signals that reveal their qualities might not be uncovered by the selection tools in place. As this is extremely hard to solve, non-discrimination should be an important part from the very first data-acquisition stage onwards.

Discriminate or accommodate

Personally, I worry about the desire to treat everybody the same. Men can also get breast cancer, but does that mean we should treat them the same as women and install a large-scale preventive screening program for men as well? The physique of people from Kenya and Ethiopia turns out to make them very fast runners. Should we then stimulate them to engage in basketball, soccer or swimming as that would be fairer? Similarly, men and women are very different but in more social behaviours like caring or risk taking, it is less clear whether this is by nature or by nurture. We would actually need to know the answer to this question in order to know whether an unequal treatment is discriminating against a group or accommodating the preferences of that group.

References

Barr, A. (2015). Google mistakenly tags black people as ‘gorillas’, showing limits of algorithms. The Wall Street Journal, 1:2015.

Dastin, J. (2018). Amazon scraps secret ai recruiting tool that showed bias against women. San Fransico, CA: Reuters. Retrieved on October, 9:2018.

Eligon, J. and Williams, T. (2015). Police program aims to pinpoint those most likely to commit crimes. New York Times, 24.

Guynn, J. (2015). Google photos labeled black people ’gorillas’. Usa Today, 1.

Kleinberg, J., Mullainathan, S., and Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. arXiv preprint arXiv:1609.05807.

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464):447–453.