ALGORITHMIC BIAS IN CURRENT HIRING PRACTICES: AN ETHICAL EXAMINATION

Dragoș BÎGU [a]*, Mihail-Valentin CERNEA [a]

[a] The Bucharest University of Economic Studies, Romania

ABSTRACT
This paper will explore the ethical consequences of using machine learning algorithms in hiring decisions, focusing on the risk of discriminating groups of people based on unjust criteria. The first section of the paper is concerned with describing the automated processes involved in current hiring practices and three sources of possible unjust discrimination: (i) the defined outputs of the algorithms involved; (ii) the way in which the predicted work performance is understood by managers; (iii) statistical correlations could be biased against certain groups of people, precluding the evaluation of individuals based on their own work performance. The second section of the paper offers a comparison between traditional cases of discrimination and this new kind of algorithmic discrimination and three solutions for mitigating the risk of discrimination in automated hiring practices, i.e., transparency, careful testing for biases that could have ingrained themselves in the software used in the hiring process, and by ensuring that the final decision is made by a human and not a machine.

KEYWORDS: *algorithmic discrimination, automated hiring practices, algorithmic bias*

1. INTRODUCTION

One of the most important innovations brought about by the Big Data Revolution is the introduction of algorithmic decision making in job recruitment around the world. While doing a lot to streamline and improve hiring processes at all levels in various companies, an ethically questionable result proves to be quite problematic for the perceived fairness of machine learning technology applied in human resources. It seems that our real-world biases, responsible for the unjust discrimination of vulnerable minorities in the job market, are making their way into algorithmic hiring decisions. This paper aims to explore the ethical questions that arise with automated recruitment practices and the risk of algorithmic discrimination that they entail. The section that follows will explore the notion of algorithmic discrimination and its possible sources. The third and final section of the paper will provide a comparison with "traditional" employment discrimination, underlining the moral benefits of algorithmic hiring, while offering some possible solutions to the ethical quandaries involved by this kind of automated processes. We will end the paper with some conclusive remarks.

2. ALGORITHMIC DISCRIMINATION IN AUTOMATED HIRING AND RECRUITMENT PRACTICES

This section of the paper is concerned with providing an account of the algorithmic discrimination that results from the large-scale use of automated hiring systems in recruitment processes around the business world. We will begin with a short description of what automated hiring actually

* Corresponding author: *dragos_bigu@yahoo.com*

consists of and then we will define the notion of algorithmic discrimination, considering the ethical challenges it poses to human resources managers. This part of the paper will end with a thorough ethical examination of the sources of algorithmic discrimination in hiring decisions.

2.1 A Very Short Introduction to Automated Hiring

As the ongoing Big Data Revolution is raging, more and more areas of human life are subjected to the application of data-centric technologies meant to optimize and assist in decision-making procedures. Automated hiring systems – digital tools meant to help employment recruiters' sort and sift through job application using extremely customizable data filters and predictive algorithms – are becoming more popular than it is those seeking jobs may believe. Most of the Global 500 companies (Barber, 2006) and all the top twenty Fortune 500 companies (Ajunwa & Green, 2019) use these kinds of recruitment tools, affecting employment at all levels of the organization, including jobs in services or retail.

It is very important to understand that this is not just a usual case of recruiters using a variety of statistical and psychological methods to select the best applicant from a pool of potential employees. That part of the hiring process is not new, as various aptitude and IQ tests have been a part of the hiring process long before Big Data was being developed. The novelty is using impersonal algorithms that have two main benefits: (i) they help recruiters profile candidates from pools of data that could not be analyzed otherwise; (ii) at least *prima facie*, they provide the recruitment process with the cold, impersonal objectivity assumed of machine learning, thus providing, in theory, better epistemic grounds for the final decision regarding who to hire for that position.

"In contrast to those traditional forms of data analysis that simply return records or summary statistics in response to a specific query, data mining attempts to locate statistical relationships in a dataset. In practical terms, it automates the process of discovering useful patterns, revealing regularities upon which subsequent decision making can rely" (Barocas & Selbst, 2016, 677). To understand the way Big Data analytics manages to bring about new patterns in the various data sets, it is useful to look at two very important concepts: *target variables* and *class labels*. The notion of target variable refers to the kind of correlations that data researchers and, ultimately, the algorithms involved are looking for in the data set, while the notion of class label divides that data in mutually exclusive categories that map all the possible values of the target variable. A simple example: say a food company needs a chef with a drivers' license – the target variable will be something along the lines of "has a driver's license" and the algorithm a recruitment firm could use would divide the pool of potential hires with cooking experience between various categories like "no driver's license", "only driver's license for motorcycles", "driver's license for automobiles", etc. This will help the food company to find exactly the chef with the driving skills it needs. Things get way more complicated the more target variables are involved in the process – the statistical work need becoming more and more inaccessible to human beings because of the complexity of the analysis necessary to draw those correlations.

These correlations alongside, sometimes, psychological tests and online applications (meant to discourage those who do not have digital competencies) eliminate, autonomously, a large part of the initial job applications, allowing few candidates to the more hands-on phase of interviews.

2.2 Unjust Discrimination in Algorithmic Hiring

As it was stated above, one of the reasons Big Data was adopted so widely in human resource management is related to the assumed objectivity of the methods involved. The problem of unjust discrimination is not new in the field of employment selection, but it has been compounded by the use of data algorithms. Usually, we can trace discrimination back to the particular unethical attitudes of the persons in charge of the hiring process, but, in the case of automated employment, the algorithms used do not intend any harm to job candidates, as they cannot intend anything at all

(statistical methods do not have wishes, moral attitudes and other features specific of human persons). Nonetheless, it has been widely recognized the even these impersonal, theoretically objective, methods do discriminate unjustly (Liu *et al.*, 2018; Baer, 2019; Winter, 2015; Ajunwa *et al.*, 2016).

One recent example is the algorithm used in setting the credit limit for the Apple Card customers. The Apple Card is a credit card service offered by the American tech giant in collaboration with Goldman Sachs. Recently, Goldman Sachs has been accused of systematically assigning low credit scores to women for no reason. The bank revealed that the algorithm used in the credit scoring system of the Apple Card would unjustly discriminate against married women because it assumed the credit history of the spouse in its determinations (Hamilton 2019).

Before going into the sources of this kind of new, unintended moral harm brought by the use of Big Data in recruitment processes it is useful to state exactly what kind of discrimination is involved in these cases. The basic idea is old: employers should not discriminate between job applications for any other reason than those related to job performance. Any other kind of discrimination is unlawful in most countries and unethical from any standpoint. The next subsection will deal with the shape and sources of unjust discrimination in algorithmic hiring practices.

2.3 Sources of Algorithmic Discrimination

In this section, we will provide a detailed account of possible sources of algorithmic discrimination. The simple fact that using an algorithm leads to members of a particular group being systematically less successful getting a job than members of some other group is not enough to conclude that the algorithm is discriminatory, as long as it is possible that members of the latter group enjoy better work performance on that position. The proof of a possible case of discrimination could only be the case in which members of the first category get hired significantly less than their actual performance would warrant. To see this issue more clearly, one could design an experiment that would show the rates of false negatives (those whose CVs are rejected despite their actual work performance) among the members of the discriminated category is much higher than average. Unfortunately, such an experiment would be tremendously difficult to put together because it all depends on the way one defines performance. This is why, as far as we know, there are no straight arguments for considering a certain algorithm to be discriminatory.

On the other hand, there are important arguments which show that hiring algorithm are not neutral or value-free. We will discuss two such arguments. First, as explained above, hiring algorithms correlate a set of variables with a target variable, which is the indicator that is predicted by the model. Any hiring algorithm needs a way to precisely define the outputs: workplace performance (as assessed by human resources department), employee retention, sales. Usually, the clients of a recruiting company define the outputs that is relevant, which may vary in accordance with organization's purposes. One of the problems is that the way in which the output is defined influences the result and can be biased. If, for instance, the soft workplace skills are considered important for the organization and, therefore, part of the output variable, women can gain advantage, compared to the situation when such skills are not taken into consideration.

Secondly, in many cases the relevant output is a measure of employees' performance and this depends on how managers assess it. In such cases, the result can preserve the bias present in employees' evaluation (Boegen & Rieke, 2018, 8). If members of a certain category are constantly unfairly evaluated, algorithm's decision will re-enforce the bias, and the applicants in that category will be constantly disadvantaged (Barocas & Selbst, 2016, 683). It is worth noting that this shortcoming of hiring algorithms cannot be necessarily found in the case of all predictive algorithms. For instance, criminal justice algorithms are used for assessing offenders' risk of recidivism, in order for courts to decide their type of sentence: prison sentence, probation, suspension, etc. In such cases, past data on offenders' recidivism are relatively neutral easy to obtain.

Algorithmic discrimination in hiring process has also a third source. As we explained above, hiring algorithms base their predictions on statistical correlations. In many cases hiring algorithms are used for screening the resumes, and, they can establish, for instance, that there is a strong correlation between the university the applicants attend and workplace performance. The correlation can be so strong that all or most applicants that attended a certain university would be rejected. This result is even more likely if we consider two facts. First, some characteristics are correlated to a high degree: for instance, applicants that attended a certain university can live a in a certain city. Secondly, given being the large number of resumes that companies receive, hiring algorithms will reject most of them before job testing or interview.[i]

Even if this correlation is real, we can ask whether it is fair to reject all applicants that attended a certain university before any testing that would prove that they are not qualified. Three remarks are useful to tackle this question. First, this form of statistical discrimination is rational for companies, since it is an easy way to manage a large number of applications that otherwise would consume much time (Boegen & Rieke, 2018, 6). At the same time, the probability to lose very good employees is not high, all the more so for the positions which do not require special abilities and for which, consequently, many applicants can be qualified. Secondly, this form of discrimination is not intentional: employers only want to find good employees and to reduce the cost of the recruitment process. Thirdly, algorithm's predictions are not the result of the personal prejudices, but are supported by data on employees' performance. In spite of this, it is unfair to treat the people according to their belonging to a certain category; they should be judged according to their individual abilities, which can be assessed only by individual tests. The decision to base hiring decisions on algorithm's predictions seems even more unfair if we take into consideration that these predictions are based on data that can be found in resumes, which are not necessarily fully relevant. The result given by a hiring algorithm can only be the best that you can get by screening resumes, but at least in some cases testing applicants can be much more relevant.

3. THE RISK OF DISCRIMINATION IN TRADITIONAL AND AUTOMATED HIRING PRACTICES

In this last section, we will make a comparison of algorithmic and traditional methods of staff selection from the point of view of discrimination risk. In many situations, traditional discrimination in the hiring process is intentional, unconscious or based on prejudices. First, some recruiters intentionally discriminate some categories of applicants. Hiring algorithms avoid for the most part this type of discrimination. Secondly, recruiters can discriminate unconsciously. For instance, studies show that candidates who are considered good-looking have an advantage. Hiring algorithms avoid this type of discrimination, since resumes do not include data about candidates' appearance. Thirdly, a large part of discrimination in the hiring process is based on prejudices, i.e. on false statistical statements about some categories of candidates. We argued above that it is not morally right for employers to base their decisions on statistical judgments about protected groups such as gender categories or ethnic minorities, even when these statistical judgments are true. However, the wrongdoing is obviously higher when these judgments are false. Given being that algorithmic hiring decision is based on correlations supported by large amounts of data, the chance of correlation being random is very low. Furthermore, algorithms can identify correlations that cannot be recognized by a human agent. Some of these correlations can work against discrimination. For instance, while a human recruiter would reject a whole category, a hiring algorithm can find that even if people in that category perform worse than average in a job, the more particular class of people that are in that category, but are graduates of a certain university perform better than average (O'Neil, 2016, 203).

These three arguments do not show that algorithmic hiring fully avoids discrimination, but that algorithmic discrimination is, from some points of view, less dangerous than traditional one.[ii]

However, from other perspective, algorithmic discrimination can be more harmful than traditional one. Thus, since hiring algorithms base their results on huge amounts of data, which can have similar effects to a greater degree than traditional hiring: with the same inputs, algorithms take the same decisions. Therefore, since traditional methods depend on many recruiters, whose opinions – biased or not – are diverse, there is a higher risk for hiring algorithms to discriminate in the same direction, their effect in society can be more pervasive.

In order to decrease the risk that hiring algorithms would be used unethically, some elements are important. First, one of the dangers of automated decision-making is that algorithms are opaque. Non-transparent algorithms cannot be assessed from an ethical point of view and improved. Algorithmic decision-making must be more transparent (Zuiderveen Borgesius, 2018, 25). HR professionals who use AI hiring tools should be informed and should understand how the software works. As far as possible, companies that use such tools should include in their ethics reports information about how they use them.

Secondly, biases should not be considered as unavoidable unintentional elements; developers and recruiters should work together to test their software in order to identify and remove the possible bias. Both parties cannot absolve from responsibility for the potentially harmful results of using AI tools. Biased recruitment software should be adjusted. One of the methods that is used for debiasing hiring software is to avoid encoding gender, race and other "protected" characteristics that can give rise to discrimination. This step is necessary in order to avoid obvious cases of discrimination, in which belonging to certain category is the only characteristic that is used for rejecting a candidate. For instance, it is possible that algorithm would rank a male candidate better than a woman candidate, even if all their other relevant characteristics are identical. A solution to avoid such cases is simply not to encode gender. But this solution is not sufficient, since other characteristics can act as proxies for protected characteristics (Zuiderveen Borgesius, 2018, 13). For instance, the postal code can be, in certain contexts, a good proxy for socioeconomic status or for race, even if these do not occur in resumes. In order to solve or at least to mitigate this problem, more complex methods should be used, but very likely with the cost of sacrificing relevant data that algorithms can use for ranking candidates.

Thirdly, hiring software should be considered only as a tool that helps recruiters, not as a final decision-maker. AI-based screening tools should be used as an initial method, which leaves a diverse and large enough pool of candidates. After the screening, phase recruiters should use other recruitment methods; for instance, at least form some positions, job knowledge tests, which objectively assess candidates' skills and abilities should play an important role.

4. CONCLUSION

As we have shown, there is a real risk that unjust discrimination will occur in algorithmic hiring, particularly when machine learning is used to predict future employee's performance in the workplace. Given that Big Data approaches to human resources are becoming common in the job market, this is a risk that should be taken very seriously. The main ethical sell point of automated decision making in job recruitment is its capacity to push aside the more traditional kind of discrimination generated by all too human biases. Algorithmic discrimination would render these kinds of practices useless, at least from an ethical standpoint.

This being said, our paper hopefully shows that the new kind of digital discrimination is not an inevitable consequence of algorithmic hiring, but can actually be avoided through transparency, careful testing for biases the could have ingrained themselves in the software used in the hiring process, and by ensuring that the final decision is made by a human and not a machine.

The risk-mitigation methods we proposed in this paper need not only apply to the field of human resources, but to any ethically challenging human activity in which algorithmic decision making is becoming standard these days. For example, unbiased decision making is key in medicine and

military operations, as the senseless loss of human life can occur if proper ethical considerations are not taken into account. Any organizational entity that uses data mining and predictive algorithms must ensure that its procedures are unbiased and non-discriminatory.

REFERENCES

Ajunwa, I., & Greene, D. (2019). Platforms at Work: Automated Hiring Platforms and Other New Intermediaries in the Organization of Work. In S.P. Vallas & A. Kovalainen (Eds.), *Work and Labor in the Digital Age*, (pp. 61-91). Bingley, UK: Emerald Publishing Limited.

Ajunwa, I., Friedler, S., Scheidegger, C. E., & Venkatasubramanian, S. (2016). Hiring by algorithm: predicting and preventing disparate impact. *Available at SSRN*. Retrieved on November 13, 2019, from http://sorelle.friedler.net/papers/SSRN-id2746078.pdf

Baer, T. (2019). *Understand, Manage, and Prevent Algorithmic Bias: A Guide for Business Users and Data Scientists*. New York: Apress.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *Calif. L. Rev.*, *104*, 671.

Bogen, M., & Rieke, A. (2018). HELP WANTED. *An Examination of Hiring Algorithms, Equity, and Bias.* Retrieved November 13, 2019, from https://apo.org.au/sites/default/files/resource-files/2018/12/apo-nid210071-1229641.pdf

Hamilton, I. A. (2019, November 12). Goldman Sachs will let people appeal their Apple Card credit limit after allegations of sexist algorithms. *Business Insider*. Retrieved November 13, 2019, from https://www.businessinsider.com/goldman-sachs-apple-card-sexism-response-2019-11

Liu, J., Li, J., Ye, F., Liu, L., Duy Le, T., & Xiong, P. (2018). An exploration of algorithmic discrimination in data and classification. *arXiv preprint: 1811.02994*. Retrieved November 13, 2019, from https://arxiv.org/abs/1811.02994.

O'neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. New York: Broadway Books.

Winter, J. (2015). Algorithmic discrimination: Big data analytics and the future of the Internet. In J. Winter & R. Ono (Eds). *The Future Internet: Alternative Visions* (pp. 125-140). Cham: Springer International Publishing.

Zuiderveen Borgesius, F. (2018). Discrimination, Artificial Intelligence, and Algorithmic Decision-Making. Strasbourg: Council of Europe. Retrieved November 13, 2019, from https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73

[i] Around 70 percent of resumes are rejected before being seen by recruiters (O'Neil, 2016, p. 114).

[ii] It should not be overlooked that algorithms are used just for screening resumes in the first phase. In the interview phase, human interviewers are prone to traditional biases, as well.