2022

Can Hiring Algorithms be Impartial Decision-Makers?

Catherine Shin

CAN HIRING ALGORITHMS BE IMPARTIAL DECISION-MAKERS?

Catherine Shin

ABSTRACT

AI decision-makers were believed to make more equitable hiring decisions than human decision-makers, who are influenced by their own internal biases. However, AI decision-makers are not inherently fairer. They are designed, trained and implemented by imperfect humans and can perpetuate the same discrimination they were supposed to combat. There is a variety of reasons for this. One reason is that the training data is biased, one group is under- or over-represented, often in a manner that reflects the existing inequalities in society. Another reason is that algorithms, on their own, can find correlations between certain traits, such as sex, and a prospective employee's chances of success at the sought-after position. Or the AI decision-makers find a proxy, a trait that is not a protected characteristic but has the same adverse impact against a group with that characteristic. This paper will focus on AI's effect on gender discrimination in hiring decisions and how to avert or rectify AI gender discrimination.

To that end, this paper will discuss recommended methods employers can utilize to prevent discriminatory outcomes, such as balancing their training data, hiding the sex of candidates or checking the fairness of the process against other AI's, fairness tools or a diverse team of experts. There will also be an examination of the legal remedies available to plaintiffs if an AI decision-maker does discriminate against them on the basis of sex, such as the disparage treatment and disparate impact theories of employment discrimination under Title VII of the Civil Rights Act of 1964 and the potential liability against the entity incorporating the AI decision-maker, if said entity is accepted as a consumer reporting agency, under the Fair Credit Reporting Act.

INTRODUCTION

It is impossible for humans to be wholly impartial decision-makers. We carry our own biases, informed by the society we were raised and live in, and those biases influence the decisions that we make. Even when we try to be objective, it is hard for us to see outside of our own preconceived notions of how the world operates and how people should behave.[1] In the hiring context, this makes it harder for human decision-makers to judge someone's fitness for a job position in a fair, even-handed manner, especially if the applicant is from a group the decision-maker is unfamiliar with or not a member of. The potential for objectivity is part of the allure behind hiring algorithms. Hiring algorithms will go through the candidates' materials and judge them solely on their qualifications and not their personal identity.[2]

However, we need to remember that algorithms do not suddenly spring into existence fully formed and operational. They have to be written, trained, and implemented by humans. It is naïve to think that just because an algorithm is not human, it cannot be influenced by human biases.[3] Algorithms have no built-in morality that tells them basing their decisions on a protected characteristic such as a candidate's sex or race is wrong. They can easily extrapolate from biased data input that historically disadvantaged groups are unattractive candidates and perpetuate the same discrimination. An example of discrimination in AI is LinkedIn's Talent March algorithm. LinkedIn is currently one of the biggest job sites on the internet. People can set up accounts with their resumes to attract potential employers. Talent March recommends candidates to employers based on the candidates the employers have demonstrated an interest in. If the employers, influenced by their own biases, show no interest in female candidates, the algorithm will not recommend female candidates to them.[4]

It is not just human bias we have to be concerned about. The algorithm can place undue importance in their predictions on characteristics that correlate closely with a target outcome but do not actually contribute to that outcome. For instance, an algorithm designed to predict cancer in patients had initially performed well during testing but had poor prediction accuracy when applied to new patients. The problem was, as part of its training, the algorithm had scanned patient's medical files, including the name of the hospital examining them. It found cancer patients tended to be treated in hospitals that specialized in cancer treatment. From that, the algorithm concluded patients treated at those hospitals must be more likely to have cancer, even though there were plenty of other medical reasons the patients could be there for. A human

[1] Kimberly A. Houser, Can AI Solve the Diversity Problem in the Tech Industry? Mitigating Noise and Bias in Employment Decision-Making, 22 Stan. Tech. L. Rev. 290, 319-20 (2019).
[2] Ignacio N. Cofone, Algorithmic Discrimination Is an Information Problem, 70 Hastings L.J. 1389, 1396 (2019).
[3] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan & Cass R. Sunstein, Discrimination in the Age of Algorithms, 10 J. of Legal Analysis 1, 5 (2019).
[4] Solon Barocas & Andrew D. Selbst, Big Data's Disparate Impact, 104 Calif. L. Rev. 671, 683 (2016).

doctor would likely not make this leap in logic but, to an algorithm, the leap made sense.[5] An example of this occurring in a hiring context is an algorithm that concluded candidates named Jared and who played lacrosse were more likely to be good employees. Luckily, the company caught this quirk before they implemented the algorithm[6] but this goes to show that employers need to remain vigilant and not just accept algorithmic outcomes blindly as unvarnished truths.

In fact, this blind faith in algorithms could pose a huge obstacle to the goal of ending employment discrimination. Believing algorithms are inherently better decision-makers leads to more employers incorporating them in their hiring process. A survey on the top 20 private employers on Fortune 500's list (which mainly consisted of retail companies) found that most of these employers used an algorithm to sort through applicants' resumes. This practice is not limited to low-wage, hourly jobs. White collar firms are also turning to hiring algorithms, with Goldman Sachs announcing in 2016 that they plan to use algorithms to search for specific keywords and experiences in applications for their summer internships and first-year analyst positions. They even took it one step further and announced their intention to let algorithms handle all their managerial decisions (hiring, firing, and promotions) in the future.[7] Hiring is becoming increasingly automatized which means, if left unchecked, algorithmic discrimination will become standardized across all industries.

The best defense against the rise of automatized discrimination is for there to be human involvement and supervision for every step of the process. From the initial stages of data collection to the final stage of actual implementation, humans need to continuously monitor the algorithm to ensure that no bias, human or algorithmic in origin, is skewing the results in a discriminatory fashion. Algorithms alone cannot be trusted to achieve this, it is only when humans and algorithms work in conjunction to compensate for each other's weaknesses that a more just and diverse workplace will come into fruition.

Part I of this paper will lay the groundwork for how hiring algorithms are trained and set-up. This background will make it easier to discuss the places in the process where biases can enter. Part II will cover the myriad of ways the algorithm can learn to discriminate. Part III proposes methods of both detecting bias in the training data as well as bias formed in the algorithms and

[5] Cofone, supra, at 1420-1421.
[6] Ifeoma Ajunwa, The Paradox of Automation as Anti-Bias Invention, 41 Cardozo L. Rev. 1671, 1689-90 (2020).
[7] Ajunwa, supra, at 1693-1694.

how to prevent the bias from affecting the final outcomes. Part IV discusses the potential liability for employers for implementing the fairness measures recommended in Part III under the disparate treatment theory of employment discrimination under Title VII but how they ultimately need not be overly concerned about it. Part V will outline the legal remedies available to applicants who experience algorithmic discrimination under the current law. The paper concludes that, despite everything that could go wrong, hiring algorithms can still fulfill their promise of ensuring more equitable hiring practices but only if humans insert themselves throughout the entire process.

I. How Hiring Algorithms Work

Employers are already implementing machine learning algorithms to make their hiring decisions. These algorithms commonly mine available data on potential job applicants from their resumes and, in some cases, their social media profiles. From this data, the algorithm predicts the likelihood the applicant will be a successful employee.[8] To understand how a hiring algorithm that is meant to be an impartial decision-maker goes on to discriminate against applicants for protected characteristics, such as sex, and where biases can potentially creep in, it would be helpful to go over how hiring algorithms are set up and trained.

a. Designing Hiring Algorithms

To set up a hiring algorithm, a human designer first must gather data. The data will be used to train the algorithm.[9] In the hiring context, the data will be collected on past employees and how well those employees performed their jobs. The assessment on past employment performance is an important component of the data because it helps the algorithm pick which features of the candidate should be considered in its predictions.[10]

The second step is for the human designer to define the target variable. The target variable is what the algorithm is being utilized to find. Employers typically make the target variable a “good worker.” But defining the measurable and observable characteristics that make an applicant a

[8] Stephanie Bornstein, Anti-discriminatory Algorithms, 70 Ala. L. Rev. 519, 531 (2018).
[9] Kleinberg et al., supra, at 22.
[10] Id.

good worker is a challenge.[11] Is it someone who speaks another language? Is it someone who has earned a college degree? Is it someone who has 2-4 years of prior experience in the field?

The third step is to pick the features the algorithm will be allowed to include in its final prediction model. At this stage, new features can be constructed (such as constructing a body mass index from data describing height and weight) and restrictions can be put in place. If the human designer wants the algorithm to ignore legally protected characteristics in its analysis, here is where they would do that.[12]

The fourth step is when the human designer steps back and the algorithm takes more control. The algorithm will identify other features to include in its final prediction model.[13] During this process, the algorithm looks for a high correlation between a potential feature and job performance. If a feature is not highly correlated to job performance, the algorithm will ignore that feature. For instance, if officers who are left- and right-handed have similar average performances, then the hiring algorithm will ignore hand dominance when looking at police officer candidates. However, if there is a noticeable difference in the average performance of past officers with college degrees and past officers without college degree (in that the former outperforms the latter), the AI will learn to prefer candidates who earn college degrees.[14]

The fifth step is to validate the algorithmic results. The human designer accomplishes this by inserting a data set into the algorithm that was not used during the training process.[15] This data set is known as the hold out set. If the algorithm computes the expected outcome from the hold out set, its results are validated and the employer can begin to use the algorithm to make hiring decisions.[16]

II. Why Hiring Algorithms Discriminate

There are many ways bias can slip into the hiring algorithm despite an employer or designer's best intentions. It could enter through the employer's subjective definition of what constitutes a good employee. The algorithm could pick it up from the training data, which is generated and collected by imperfect humans and tends to reflect an unequal society. Or the algorithm could,

[11] Id.
[12] Id. at 23.
[13] Id.
[14] Id.
[15] Id. at 24
[16] Id. at 24-5.

on its own, select and stress the importance of features that do not actually reveal much about the applicant's job relevant skills but correlates closely with a protected characteristic such as sex. In the end, that feature has the same discriminatory impact as the algorithm basing its prediction on the protected characteristic. We also cannot ignore the possibility that employers will take advantage of these hiring algorithms' idiosyncrasies to hide or "mask" their intentional discrimination.

a. Biased Target Variable

The target variable is what the algorithm seeks to find out. In the hiring context, the target variable is the candidate's potential to be a good employee. It is the human designer who defines what that term means and how to measure it.[17] The definition can be formulated based on observable characteristics of the candidates, such as their ability to complete a certain task within a set time period.[18] Barring certain protected characteristics such as race or sex, the human designer can decide which observable characteristics to include in their definition of a good employee, even though their judgement may be clouded by their own biases.

Even if the designers are not biased against protected groups, they could still incorporate the bias into the target variable's definition by including subjective characteristics and managerial decisions that reflect past discrimination. For example, the human designer could decide to use past performance review scores to determine who constitutes a good employee.[19] Initially, this seems like a good plan. Performance reviews are meant to measure how well an employee does at their job. However, those performance review were likely scored by a human evaluator. Intentionally or unintentionally, a biased evaluator could have skewed the results. A sexist grader, for instance, could have penalized the female employees for showing aggression while praising male employees for exhibiting that same trait.[20] The human programmer would therefore have introduced bias into the target variable and, subsequently, the algorithm.

b. Biased Training Data Set

[17] Jason R. Bent, Is Algorithmic Affirmative Action Legal?, 108 Geo. L.J. 803, 811 (2020).
[18] Barocas et al., supra, at 679.
[19] Id. at 679-80.
[20] Bent, supra, at 812.

It is often said that an algorithm is only as good as its training data. Garbage in, garbage out. There are two major ways that 'garbage' training data can affect an algorithm. One is if the training data reflects past discrimination. The algorithm trained on that data will reproduce that discrimination.[21] For instance, St. George's Hospital implemented an algorithm to help them sort through medical school applicants. However, the administration had a history of discrimination against racial minorities and female candidates. This resulted in the admissions algorithm discriminating against members of those groups who were otherwise qualified for admission.[22]

The other way a training data set can compromise the fairness of an algorithm is if certain populations are either over- or under-represented in the data. Both under-representation and over-representation in the data can lead to discriminatory results for protected classes. Amazon advanced algorithmic discrimination as a result of a population being under-represented in the data. Amazon trusted their algorithm to rank job applicants impartially. The designers trained the algorithm with data collected on successful employees. The employees were overwhelming male. The designers failed to realize this meant their training data over-represented the male population and under-represented the female population. Because of the bias in the training data, Amazon's algorithm associated successful candidates with maleness. The algorithn not only rated female candidates lower but discarded any resumes with the word "women's."[23] Additionally, if there is misinformation recorded about a member of an under-represented population in the training data, then data pertaining to that group will be impacted at a higher rate than it would be if the individual was a part of a well-represented population.[24]

Over-representation in the data can be dangerous for protected classes as well. If past employers were prejudiced against members of a protected class, the employers would have been more likely to closely monitor those employees or notice and report their mistakes. As a result, the algorithm would get a skewed idea as to how well members of a protected class performed on the job.[25]

c. Feature Selection and Proxies

[21] Barocas et al., supra, at 682.
[22] Id.
[23] Houser, supra, at 334-35.
[24] Barocas et al., supra, at 684.
[25] Id. at 687.

An algorithm (or a human designer[26]) selects features it considers helpful in predicting which candidates will make good future employees. Unfortunately, certain features can lead to less favorable outcomes for protected classes. For instance, it is a common practice for hiring algorithms to consider a candidate's academic credentials. As they do this, the algorithm attaches a lot of significance to the prestige of the college the candidate attended, despite the fact the prestige of the institution will likely not reveal much relevant information about the applicant's capabilities. And some of these prestigious colleges have a history of discrimination against women and racial minorities. This results in otherwise qualified candidates belonging to protected groups being denied job opportunities because of erroneous feature selection.[27]

Additionally, if a feature is highly correlated to a protected characteristic, then that feature could become a proxy for the characteristic. If the algorithm discriminates against protected class members on the basis of that proxy, it would have the same adverse impact than if the algorithm discriminated against them on the basis of a protected characteristic.[28] For example, job tenure can constitute a proxy for sex because women are more likely to quit their jobs to take care of their families.[29] Even if a human designer coded the algorithm to ignore sex, if job tenure was selected as a feature then the algorithm would still disproportionately exclude women. It is because of proxies that simply blinding the algorithm to protected characteristics will not cleanse the algorithm of bias.[30]

It is difficult to predict which characteristics act as proxies. What were not proxies in the beginning can become proxies later on.[31] Or it is only a combination of the characteristics aggregated together that results in the proxy, so eliminating one characteristic will not eliminate the proxy.[32] Even if it were possible to detect all proxies, blocking them might sacrifice accuracy in the predictions because the proxies could be providing useful information on the applicant's relevant skills.[33] For instance, someone's education level can be related to job performance but it can also be closely tied to the applicant's race. Developers must decide how they will balance fairness against accuracy.

[26] Id. at 688.
[27] Id. at 689.
[28] Id. at 691.
[29] Bent, supra, at 813.
[30] Id.
[31] Cofone, supra, at 1413-14.
[32] Id. at 1413.
[33] Id. at 1414.

d. Masking

It is always possible that an employer will manipulate an algorithm to hide their intentional discrimination.[34] This is known as masking. All ways algorithms can unintentionally learn to discriminate against protected classes can just as easily be applied intentionally. The employers or designers could knowingly collect training data that reflects past discrimination or represents protected groups in a harmful way.[35] Employers can even identify the proxies the algorithm uses for class membership and discriminate against the applicants the proxies single out without having to look up the candidate's identity.[36] Employers can then pass the responsibility for the discriminatory results onto the algorithm.

III. Solution to Algorithmic Discrimination

Despite these issues, hiring algorithms can still be used to produce more equitable hiring outcomes. Humans just have to be committed to monitoring and correcting the algorithms throughout their development, training, and implementation. Data, for instance, needs to be made as bias-free as possible, preferably before it is introduced to the algorithm. In some cases, this requires manipulating the data even though it accurately reflects reality, otherwise the algorithm would just be repeating the same discriminatory hiring choices as past human decision-makers. Proxies formed in the algorithm should either be removed, transfer proxies introduced to spread the burden away from the protected classes, or applicant's membership to protected groups hidden so the algorithm does not judge them based on their personal identity. In addition, there exists many test methods that have proven successful in measuring the algorithm's fairness and what factors goes into its final predications. Therefore, as long as we remain mindful of hiring algorithms' potential problems and invest the time and resources to keep them as accurate and fair as possible, nothing should stop employers from taking full advantage of this modern technology.

a. Balancing Training Data Sets

[34] Barocas et al., supra, at 693.

[35] Id. at 692.

[36] Id. at 692-3.

Biased training data can lead to biased algorithmic decision-making. Therefore, it is crucial to ensure that the training data is balanced.[37]

One solution to detect and eliminate bias in training data sets is to implement other machine-learning algorithms to test for it before the data is used for training. Accenture and MIT's Computing Science and Artificial Intelligence Laboratory have both developed AI's for this exact purpose.[38]

Besides other algorithms, a diverse team of algorithm developers retained from the start of the project increases the probability that an issue with the training data will be spotted earlier.[39] A diverse team could have caught the problems in St. George's or Amazon's data sets before the algorithm was trained, implemented, and discriminated, opening the institutions up to legal liability and damaging their reputations. Employers should therefore aim to hire diversely in algorithmic development because, even with good intentions, developers are only human and their biases will inevitably impact their work.

If a protected group is under-represented in the training data, then a human developer can drastically increase the diversity of the data points reviewed to compensate. For instance, ZestFinance, a credit reporting company, uses an algorithm that considers tens of thousands of pieces of information beyond what is customarily reviewed if they need to create a credit score for an individual who has little to no credit history. The same could be done with training data. If women are under-represented in the training data, then the algorithm should increase the diversity of data points considered for female applicants.[40]

If a protected group is over-represented in a harmful manner in the training data, the text classifier model could correct the resulting bias in the algorithm. In a text classifier model, wrongful negative associations are balanced out with neutral associations. This method has proven successful before. It stopped an algorithm designed to flag toxic comments on Wikipedia Talk Pages from falsely labeling every comment that included the word gay as toxic regardless of its context. The reason the algorithm gave so many false positives was because it had observed many instances where the word gay was used negatively. To counteract this, new data points were introduced that mentioned gay in a neutral way, such as saying, "I am gay," or "I am

[37] Houser, supra, at 335.
[38] Id. at 338.
[39] Id. at 339.
[40] Id. at 337.

a gay person."[41] This method mitigated the training data bias as well as increased the effectiveness of the algorithm. This method could also work to balance out negative stereotypes or references to protected groups in training data sets.

As the above example demonstrates, sometimes it is necessary to skew the training data to reflect a better society than the one we currently inhabit.[42] Marginalized populations have not had the same access as the rest of the population to education, training, or mentorships. As a result, members of these groups might perform worst, on average, in certain jobs than non-protected populations. This is known as differential observability.[43] Therefore, these populations are likely to be under-represented in training data sets and the data that is collected on them would lead to the algorithm discriminating against them. Therefore, to obtain equitable results, "boosting" and "reducing" groups should be used to manipulate the data for a fairer outcome. If one group is over-represented in the data, the developer can discard results from this group. This is "reduction." If another group is under-represented, the developer can duplicate results from this category. This is "boosting."[44] For instance, if male employees are over-represented in the training data, the developer can discard results from past male employees and duplicate results from past female employees and prevent the algorithm from developing the same biases as society at large.

b. AI Data Transparency Model

If employers formulate a regulatory scheme for their hiring algorithms, it would improve the quality of their training data and decrease the probability of discriminatory outcomes. There are four main components to creating such a plan, known as an AI Data Transparency Model. The model's guidelines are general so there will be original suggestions on how to specifically tailor the model to hiring algorithms.

For the first component, the employer should regularly audit the data their algorithms are exposed to.[45] Audits should focus on the sources of the data, the data's content and whether the

[41] Id. at 342.
[42] Cofone, supra, at 1422.
[43] Id. at 1420.
[44] Houser, supra, at 336.
[45] Shlomit Yanisky-Ravid & Sean K, Hallisey, Equality and Privacy by Design: A New Model of Artificial Intelligence Data Transparency via Auditing, Certification, and Safe Harbor Regimes, 46 Fordham Urb. L.J. 428, 474 (2019).

data's use complies with any regulatory limits, such as federal laws prohibiting employment decisions based on a protected characteristic like sex.[46] For the second component, the human developers must save and store the training data, in case an issue crops up in the future and the data needs to be re-examined.[47] For the third component, employers should hire third parties who did not help design the algorithm, preferably individuals from diverse backgrounds, to audit the data.[48] The audit should have established standards that strike a balance between ensuring the data is unbiased against protected groups and acknowledgement there is no neutral data because every facet of a person's life is impacted by their identity, including their sex and race.[49] Sometimes, this requires that the training data be manipulated to reflect a better society. Sometimes, this means keeping a feature that is closely tied to a protected characteristic but also to the target variable, such as level of education. Once the audit is finished, the third parties should be the ones to certify the results. The fourth component involves a change in the law. It calls for a safe harbor provision for the employers that did not purposefully or negligently collect biased training data.[50] Despite people's best efforts and intentions, algorithms will sometimes run amok. If employers are held liable for algorithmic discrimination regardless of what preventative measures they took, they will have no incentive to change. However, even if the law never officially provides a safe harbor for employers, compliance with the rest of the regulatory scheme will help them mount a strong affirmative defense in an employment discrimination suit.

Adopting this model will make it easier for employers, if problems arise, to identity what went wrong, whether it was biases from the training data, the developer, or the algorithm itself, and what can be done to fix it.[51] It is also recommended that the level of transparency aimed for should depend on the algorithm's use.[52] For hiring algorithms, transparency should require developers to have a basic understanding of how the algorithm operates, so they can explain how the algorithm was designed, the data used to train it and its success at predicting good candidates. To measure the algorithm's success rate, there must be a record kept of both false positives

[46] Id. at 473.
[47] Id. at 474.
[48] Id.
[49] Barocas et al., supra, at 721.
[50] Yanisky-Ravid et al., supra, at 476-77.
[51] Id. at 477-78.
[52] Id. at 478.

(applicants predicted to be good employees but were not) and false negatives (applicants predicted to be bad employees but were not).

Transparency measures might raise initial costs during the training stage, but the trade-off is that otherwise qualified employees will not be screened out even though they might be the best fit for the job. And, perhaps most importantly to the employer, the risk of legal liability for employment discrimination and reputational harm to their organizations resulting from publicized instances of algorithmic discrimination will decrease.[53]

c. Dealing with Proxies

It is possible to blind an algorithm so it does not "see" sex or race. But algorithms excel at finding proxies, it is what they are designed to do, and it will find features that closely correlate to protected group status. The applicant's sex or race still ends up affecting the algorithm's final prediction.[54]

Eliminating the proxies may be beneficial to the algorithm because the proxy only provides useful information on predicting which candidate fits the human developer's definition of a good employee, and that definition is not always helpful or accurate.[55] And even if eliminating the proxy makes the algorithm less accurate, the loss in accuracy should be worth it to avoid perpetuating discrimination and legal liability for said discrimination.[56]

Besides eliminating proxies (which may not always be feasible), one proposed solution for dealing with proxies is to shift discrimination from one population to another through transfer proxies. Transfer proxies target a group that has some overlap with the protected class. By reducing proxies, the algorithm developer can hone-in on a specific target group.[57] In this scenario, the members of the protected class who are not members of the target group are not discriminated against, members of both the protected group and the target group experience the same level of discrimination, and members of the target group experience discrimination they may not have experienced otherwise.[58] This method fails if the targeted group is another

[53] Id. at 479.
[54] Cofone, supra, at 1412-1413.
[55] Id. at 1441.
[56] Id. at 1441-42.
[57] Id. at 1418.
[58] Id. at 1416.

protected group or historically disadvantaged minority.[59] For instance, targeting trans applicants instead of female applicants means that the trans applicants suffer from more employment discrimination and trans female candidates experience the worst of both worlds as they belong to both groups. In such case, the employer has opened themselves up to a plethora of potential employment discrimination lawsuits.

However, proxies can also be expanded to distribute the discrimination across a large population (ideally, the whole population) and avoid targeting one group directly. This may be preferable to reducing proxies because it comes closer to equalizing the playing field.[60]

Another alternative to blinding the algorithm to protected characteristics is to gather information on the protected classes while simultaneously obscuring an individual applicant's protected class-based status. This is actionable privacy. This can be accomplished through multi-party computation. The applicant encrypts their personal data before they send in their application and then the company, assisted by a regulatory body, encrypts the applicant's data. This way no one knows all the data all at one time, including who belongs to which groups.[61] The algorithm could also do this task on its own, collecting data on groups while encoding the sensitive protected characteristics so it is not aware of which applicants are members of the protected group.[62] Through actionable privacy, important data is collected without unfairly impacting the applicants.

d. Fairness Tests

People describe AI as a "black box" because it is difficult to identify why an algorithmic decision-maker came to the decision that it did. This makes adjusting a biased algorithm a challenging task. However, there are methods for checking if the algorithms are producing fair outcomes.[63]

One method is called "group fairness." This involves sorting candidates into groups based on a protected characteristic.[64] If the protected characteristic was sex, then male and female applicants could be sorted into two groups. Then an evaluator will compare the number of

[59] Id. at 1418-19.
[60] Id. at 1419.
[61] Id. at 1425.
[62] Id. at 1426.
[63] Houser, supra, at 340.
[64] Bent, supra, at 817.

positive outcomes for each group. If male group members get positive outcomes (i.e. the algorithm saying they will be good employees) at a higher rate than female group members, or vice versa, that is a sign the algorithm's decision-making process is being influenced by the applicant's protected characteristic.

To promote group fairness, it is not necessary for the groups' results to match up exactly. It is instead recommended that employers adopt the Equal Employment Opportunity Commission's four-fifth's rule of thumb for comparing differences in selection rates produced by an employment practice. Ideally, the employer should strive for a positive outcome prediction rate for the lower rated group that is no less than four-fifths of the positive outcome prediction rate for the higher rated group.[65] This prevents the remedy from becoming a quota.

In addition to "group fairness," there is "individual fairness." For this test, two candidates who possess similar qualifications (e.g. graduated from the same college, interned at the same company, have the same references, etc.) are selected. The only significant difference between the two is one belongs to a protected class and the other does not. The algorithm should give them both the same result.[66] If the algorithm approves of only one of them, then that is a sign the algorithm is being influenced by that protected characteristic.

It is recommended, however, that employers not utilize both individual fairness and group fairness at the same time. This is because what would be fair for the group may not be fair for the individual.[67] If the algorithm predicts that 20 male candidates would be good future employees, then group fairness requires that at least 16 female candidates be declared good future employees. Even if this leads to a female candidate getting a positive outcome when a male candidate with similar qualifications would not.

Other methods for examining algorithm fairness include quantitative input influence, adversarial debiasing, counterfactual testing, and hard debiasing. Quantitative input influence measures the effect a data point has on the algorithm's final predictions. The more influential the data input, the greater impact it has on the final prediction.[68] This is helpful because if a protected class-based characteristic or its proxy has a great impact on the final prediction, then the employer is alerted that the algorithm needs adjusting. Adversarial debiasing is an algorithm

[65] Id. at 817-18.
[66] Id. at 819.
[67] Cofone, supra, at 1434.
[68] Houser, supra, at 341.

designed to test whether the hiring algorithm is predicting which applicants will be good employees from data input, like the applicant's work history or level of education, without predicting the candidate's protected characteristic. If the adversarial debiasing algorithm comes to the same conclusions as the hiring algorithm but is unable to predict whether the applicant is part of a protected group, then the hiring algorithm is a fair decision-maker.[69] Counterfactual testing is already used to test fairness in law school admissions. In counterfactual testing, an algorithm is fair if it gives the same prediction for the target variable regardless of whether the individual is a part of a protected group. This means an applicant can enter their information first under their real sex, then as another, and the results would be the same both times.[70] However, for counterfactual testing to work, an analyst needs to create a casual model that demonstrates their theory on how the sensitive variable (e.g. sex or race) interacts with the target variable and other observed variables. Hard debiasing, unlike the rest of the tests, is a method that specifically targets sex discrimination in algorithms. Researchers at the 2016 Neural Informal Processing Systems (NIPS) conference have shown the system can detect and remove gender stereotypes in the hiring algorithm that came from biased training data.[71]

And, as with training data, AI can be utilized to search for bias in the hiring algorithms. Major companies have developed resources to monitor an algorithm's quality. IBM's AI Fairness 360 has an open source library available online which includes 70 fairness metrics and ten advanced bias mitigation algorithms.[72] Other tools include Facebook's Fairness Flow, Pymetrics' open-source Audit AI Tool, Accenture's Toolkit, and Google's What-if Tool.[73]

With all these techniques for examining and correcting biases in hiring algorithms, it might actually be easier to peek inside of an algorithmic decision-maker than it ever would a human one.

IV. Potential Legal Liability For Proposed Solutions

Employers need to take protected class-based characteristics into consideration in order to successfully root out biases in training data sets and the algorithms themselves. This leads to a

[69] Id. at 342-43.
[70] Id. at 342.
[71] Id. at 341.
[72] Id. at 343.
[73] Id. at 343-44.

(not unfounded) concern on the part of employers that implementing such protected class-based conscious measures will open them up to liability from rejected applicants under a disparate treatment theory of employment discrimination.[74] Candidates will allege that they were adversely affected by the employer's impermissible consideration of protected characteristics in their decision-making process, even if the inclusion was to make the outcomes fairer and less arbitrary. However, employers should not worry themselves. Valid affirmative action plans have been accepted by the courts as legitimate non-discriminatory reasons for hiring decisions. The methods suggested in Part III will fit into the legal definition carved out for valid affirmative action plans and act as a defense in those disparate treatment claims.

a. Liability Under the Disparate Treatment Theory

In Ricci v. DeStefano, the Supreme Court held that the (mostly white) firemen who were denied promotions when the city of New Haven refused to certify test results because a disproportionate number of black and Latino firemen failed the test[75] had a valid disparate treatment claim against the city.[76] The Court found the city lacked a valid non-discriminatory reason to discard the test results since their whole reason for doing so was a reaction to the racial make-up of the firemen who passed and failed the test.[77] The concern is that under Ricci's standard, plaintiffs could sue for employment discrimination under a disparage treatment theory, alleging the algorithm rejected them because it impermissibly relied on a protected characteristic such as their sex or race.

However, the Supreme Court also established in Johnson v. Transportation Agency that a valid affirmative action plan is an acceptable non-discriminatory reason for failure to hire a candidate.[78] The Second Circuit has held an affirmative action plan is a forward-facing procedure aimed at ensuring equal opportunity and eradicating future discrimination.[79] New Haven's conduct was backwards-facing. They conducted the tests, the results came back, they realized the tests lead to an inequitable result and tried to "fix" the result after the fact. The proposed solutions in this paper are focused on ensuing algorithmic discrimination never occurs in the first

[74] Bent, supra, at 808.
[75] Id. at 827.
[76] Id.
[77] Id. at 827-28.
[78] Id. at 835.
[79] Id. at 836-37.

place or catching it before it affects real people. The proposed solutions to de-bias training data sets and algorithms would therefore constitute affirmative action plans.

As for what constitutes a valid affirmative action plan, Weber-Johnson gives three factors to consider. The first factor is to point to an existing imbalance in workforce demographics in traditionally segregated job categories.[80] Amazon, if they had utilized the recommended methods to ensure their training data was unbiased, would have no issue proving the tech sector has traditionally been male dominant to justify said methods.[81] The second factor requires that the plan not unnecessarily trample others' rights or pose an absolute bar to their advancement.[82] The proposed solutions in this paper are simply to ensure a more even playing field for applicants, regardless of their protected characteristics. The purpose is not to force employers to hire a set amount of applicants of a certain sex.[83] On a related note, the third factor holds that any affirmative action plan with a set quota or ratio be a temporary measure.[84] However, none of the proposed solutions create set quotas or ratios.[85] Even group fairness, which measures how often members of one group get a good algorithmic prediction compared to the members of a protected group, does not require that the number of candidates of each group to receive a positive outcome be the same. It just recommends employers follow the four-fifths rule of thumb laid out by the EEOC. Therefore, an employer should not fear disparate treatment lawsuits for considering protected characteristics in bias prevention measures and, in fact, should take advantage of those measures to ensure a fair hiring algorithm.

V. The Possible Legal Remedies For Algorithmic Employment Discrimination

Legal liability both incentivizes employers to keep their algorithms bias-free[86] and allows applicants to recover if the algorithm does end up discriminating against them. While the algorithm is a relatively new addition to the hiring process, employment practices discriminating against applicants due to a protected characteristic is not a new phenomenon, and the law offers remedies for it. One major source of employment discrimination law is the disparate treatment

[80] Id. at 837.
[81] Id. at 837-38.
[82] Id. at 837.
[83] Id. at 840.
[84] Id. at 837.
[85] Id. at 840.
[86] Bornstein, supra, at 537.

theory (specifically the anti-stereotyping theory) and the disparate impact theory under Title VII of the Civil Rights Act of 1964. While both theories have their advantages and disadvantages with the added complication of machine learning algorithms, the applicant could still make a strong case for themselves under either theory. Or, the applicant can opt instead to go after the algorithm designer (which could also be the employer) or the company who offered their algorithm for the employer's use. The reason for this is those entities can constitute consumer reporting agencies and therefore fall under the purview of the Fair Credit Reporting Act, which penalizes agencies generating inaccurate consumer reports.

Therefore, even if the discrimination comes from an algorithm, applicants are still left with legal recourse for that discrimination, which serves as an advantage for society as a whole because it keeps the employer from getting too complacent about their algorithm.

a. Civil Rights Act of 1964

Under Title VII of the Civil Rights Act of 1964, a plaintiff can file a suit against an employer that failed or refused to hire them because of their race, color, religion, sex or national origin.[87] There are two theories under Title VII for employment discrimination. One theory is disparate treatment.[88] Disparate treatment originates from the language of the statute that employers are prohibited from taking any adverse action against an employee or job applicant "because of" a protected characteristic.[89] The plaintiff, as part of their burden, has to show there was discriminatory intent on the part of the employer.[90] Because it is unlikely that an employer would outright admit they did not hire a job applicant because of a protected characteristic, the courts accept circumstantial evidence of discriminatory intent.[91] In the context of algorithms however, even allowing for circumstantial evidence of discriminatory intent, the plaintiff has a difficult burden to overcome.

For one, algorithms can learn to be discriminatory on their own or because of the carelessness, but not the intent, of the one who trained them. As the now infamous Amazon story goes, the company wanted their machine learning algorithm to rank job candidates impartially.

[87] Matthew U. Scherer, Allan G. King & Marko J. Mrkonich, Applying Old Rules to New Tools: Employment Discrimination Law in the Age of Algorithms, 71 S.C.L. Rev. 449, 457-58 (2019).
[88] Id. at 458.
[89] Id. at 459.
[90] Id.
[91] Id.

They just failed to realize in time that their training data skewed heavily male. This was not a result of deliberate discriminatory intent on their part. They just did not consider how the biases baked into their training data set (which reflected the societal biases at the company at large) would affect the machine learning algorithm. For a plaintiff suing under disparate treatment theory, some legal scholars believe unless and until the plaintiff could show the employer deliberately sabotaged the algorithm, it would be hard for them to prove the employer intentionally discriminated against them. Therefore, while disparate treatment theory would be helpful in cases where the algorithm was used to mask the employer's intentional discrimination, it would be much harder to prevail on a claim in cases where there was no human intentionally skewing the algorithmic outcomes.

However, an individual could still overcome this hurdle and make a successful disparate treatment claim under an anti-stereotyping theory. Disparate treatment embodies the anti-classification approach to discrimination. It holds that everyone should be treated exactly the same, ignoring how past and current discrimination continues to impact minorities and women.[92] While this principle can even bar employers who want to take protected characteristics into consideration to benefit marginalized groups, it also requires that an individual be treated as an individual. They should not be judged based on class-based stereotypes[93], even if there is a kernel of truth in the stereotype.[94] For instance, in Los Angeles Department of Water & Power v. Manhart, the Supreme Court held that the employer could not force female employees to contribute more to the pension fund than their male counterparts based on the stereotype that women live longer on average than men. The employer assuming all of their female employees would fit this stereotype and penalizing them based off that assumption was a violation of Title VII under the disparate treatment theory.[95]

It was in Price Waterhouse v. Hopkin where the Court first articulated the "stereotype theory" of liability. Plaintiff alleged her employers passed her over for promotion at the firm due to her sex. She was an exemplary employee but her employers criticized her for acting too "macho" and "masculine." They told her she should dress and act more femininely if she wanted

[92] Bornstein, supra, at 541.
[93] Id. at 544.
[94] Id. at 547.
[95] Id.

to be a partner. She won her case under disparate treatment theory.[96] The case showed that evidence of stereotyping is enough to prove employers possessed a class-based motive. The plaintiff then does not need to offer comparator evidence (evidence that similarly situated individuals outside of the protected class were treated better) which some courts require in a disparate treatment claim.[97] And stereotyping, as a practice, also suffices as evidence of intent.[98]

Intent does not require the employer to consciously realize they are discriminating against protected classes. Intent can encompass conduct where the employer intentionally applies the same subjective employment practice to all employees or prospective employees, and the practice negatively impacts members of a protected class. [99] For instance, imagine that an employer always runs an applicant's resume through a hiring algorithm and makes hiring decisions based on the algorithm's predictions. The algorithm's training data set was based on past employee records. However, unbeknownst to the employer, the records were generated by a long-time manager who monitored female employees much more closely than their male counterparts because he believed women wasted too much time gossiping with each other. As a result, he reported female employees for minor infractions at a higher rate than male employees. The past discrimination in those evaluations is transferred to the algorithm. The algorithm then tosses out any resume with a traditionally feminine-sounding name to screen out female applicants. The algorithm is arguably no longer a facially neutral employment practice so disparate impact theory would no longer apply. But, under disparage treatment theory, this would constitute discriminatory intent as well as implicate the anti-stereotyping theory. The burden would then shift to the employer to show there was a legitimate non-discriminatory reason behind their hiring decision.[100] In a situation like this, where the algorithm is influenced by gender stereotypes, it would be difficult for employers to meet their burden.

The other theory under Title VII is disparate impact which was established in Griggs v. Duke Power Co. In Griggs, black employees alleged in a class action suit that their employer's policy requiring that new hires, in all but the lowest paying departments, either possess a high school diploma or pass a general intelligence test constituted employment discrimination. Both of these

[96] Id. at 547-48.
[97] Id. at 549-50.
[98] Id. at 559.
[99] Id. at 558-59.
[100] Id. at 560.

criteria disproportionately affected black candidates.[101] The Court of Appeals held the requirements did not violate Title VII because the tests made no explicit distinction between employees based on their race and they found no discriminatory intent. The Supreme Court reversed[102], holding that Title VII prohibited practices that were fair in form but discriminatory in practice.[103] Because of this, most scholars believe disparate impact theory is the better fit for a plaintiff alleging algorithmic discrimination. The AI is meant to be an impartial decision-maker ('fair in form') but without careful training, monitoring and implementation, the AI can transform into a tool of discrimination instead ('discriminatory in practice').

Rather than trying to discern the intent behind an algorithm's decisions (if intent even exists in modern-day AI) the plaintiff can point to the algorithm's predictions as evidence instead. If, for instance, a 'significant' number of candidates that share the same sex are disproportionately impacted by the hiring algorithm that can go to proving disparate impact.[104] The problem is determining what significant means in this context, whether it is in the statistical sense or the colloquial sense.[105] The Court in Ricci v. DeStefano suggested that statistical significance is preferred.[106] This is good for the plaintiff because the larger the pool of candidates the algorithms are feed, the greater the likelihood the number of candidates rejected belonging to a protected class will be statistically significant[107] compared to the proper labor market. [108]

However, even if the plaintiff provides statistically significant evidence that the algorithm was "fair in form but discriminatory in practice," the employer could still raise the business necessity defense to avoid liability. The test articulated in Title VII of the Civil Rights Act of 1991 for business necessity defense states that the practice must be "job related for the position in question and consistent with business necessity."[109] For this defense, courts have required employers to demonstrate a connection between their selection procedure and specific key aspects of job performance, a process known as validation.[110]

[101] Scherer et al., supra, at 460.
[102] Id. at 460-61.
[103] Id. at 461.
[104] Id. at 462.
[105] Id.
[106] Id. at 464.
[107] Id.
[108] Bornstein, supra, at 554.
[109] Scherer et al., supra, at 464.
[110] Id. at 469.

This could be a plaintiff-friendly standard in the algorithmic decision-making context because an algorithm finds patterns and can give weight to characteristics without input or direction from the human employer (if the employer was even involved in the first place). For instance, Amazon's hiring algorithm decided on its own that maleness was a desirable trait in a candidate. The system relied on an impermissible characteristic to make its prediction and Amazon could not then argue that a candidate's sex is related to specific key aspects of job performance. Perhaps this argument would work if the job involved manual labor since, in the context of 14th Amendment equal protection, the Court has allowed parties to take the physical differences between the sexes into consideration, but it would be a weak argument in any other context.

However, one challenge that may crop up for the plaintiff is if the algorithm found a proxy that had the effect of disproportionately singling out candidates based on their sex. Job tenure can act as such a proxy because women are the ones who typically quit work to fulfill family obligations. But the proxy is not explicitly tied to femaleness, and large gaps in a candidate's work history could indicate they are unfit for the job.[111] Another challenge is that everything a hiring algorithm does is inherently job related. Its main function is to use training data to determine the criteria for a desirable employee and then select employees based on that criteria. The algorithm would therefore be self-validating[112], at least under a "criterion validation" study. A study that courts have accepted for business necessity defenses.[113]

According to Albemarle Paper Co. v. Moody, if an employer successfully raises the business necessity defense, the plaintiff has to show that other tests or selection methods would have served the employer's legitimate interest in 'efficient and trustworthy workmanship,' without the same discriminatory impact[114], and the employer refused to adopt them.[115] There exists some confusion over how to show an employer failed to adopt a less discriminatory practice. Case law though indicates that a plaintiff who shows the defendant could have made their own employment practice fairer, and did not, would suffice.[116] As for the first prong, plaintiffs could introduce any and all the methods already outlined in this paper, such as hiring objective third

[111] Barocas et al., supra, at 721
[112] Bornstein, supra, at 555.
[113] Id.
[114] Scherer et al., supra, at 471.
[115] Bornstein, supra, at 557.
[116] Id.

parties to audit the training data sets or using transfer proxies to spread the bias across a large population.

Concerns have been raised that even if plaintiffs did point to these methods of debiasing algorithms, they still would not meet their burden because the methods take money and time to implement.[117] Because of their costs, employers could argue they are not reasonable alternatives. However, while this may be true for some of the methods, it is not true for all of them.[118] There are methods that are not prohibitively expensive or time consuming. IBM, for example, has an open source library offering cutting edge algorithm bias detecting tools. The plaintiff can highlight those methods to show the employer could have improved its own algorithm so it would function and pick candidates in a more equitable way, but the employer failed to, and win on their disparate impact claim.

b. Fair Credit Reporting Act

Algorithms can serve many different but related functions in hiring. They can gleam information on an applicant from their resume or social media presence, update an applicant's profile or direct employers to the best applicants based on the qualifications the employer lists as desirable or past applicants the employer has looked up. These functions give plaintiffs a possible avenue for relief or a way to build their disparate treatment claim against said employers under the Fair Credit Reporting Act.

The Fair Credit Reporting Act regulates consumer reporting agencies.[119] An entity can be classified as a consumer reporting agency (even if they claim not to be one[120]) if it gathers information on consumers for the purposes of furnishing a consumer report.[121] A consumer report is any form of communication that describes a consumer's "…character, general reputation, personal characteristics, mode of living which is used or expected to be used…as a factor in establishing the consumer's eligibility for…credit, insurance…or employment purposes" for the benefit of third parties.[122] Algorithms fit the definition of a consumer reporting agency under the statute because the algorithm's job is to generate predictions about a

[117] Id.
[118] Id.
[119] Ifeoma Ajunwa, The Paradox of Automation as Anti-Bias Intervention, 41 Cardozo L. Rev. 1671, 1736 (2020).
[120] Id. at 1739.
[121] Id. at 1736-37.
[122] Id. at 1737.

candidate's fitness for a position, which goes to reporting on their character or personal characteristics for the purpose of employment. If employers reject a candidate because they relied upon the report from an algorithm, the employer has an obligation under the act to inform the adversely affected person of said reliance.[123] The rejected applicant, as the consumer, has a statutory right to request in writing disclosure of the nature and substance of the information in their file at the time of their request. Employers have 5 business days to comply.[124] If an algorithm compiled the consumer report, employers should, at the very least, disclose they used an algorithm, the data on the applicant that was inputted, the specific data points the algorithm focused in on, and the algorithm's final recommendation. Legal scholar Ifeoma Ajunwa proposes this provision could give the plaintiff evidence for their disparate impact claim[125] but I argue it can provide evidence for a disparate treatment claim as well. The reasons behind the algorithm's recommendation may not be fully clear from this information alone but it could assist in showing that the employer relied on an algorithm that considered a protected characteristic or a proxy in its analysis.[126] For example, if in the past, employers overwhelmingly looked at male applicants, it could be shown the algorithm looked up the sex of the applicant to complete its prediction. The report might then help the applicant build their case for a disparate treatment claim under Title VII, showing they were denied employment because of a protected characteristic.

However, the plaintiff also has the option, in cases where the algorithm takes a more active role in directing employers to applicants and modifying applicant's profiles based on publicly available information, to allege negligence against the algorithm developers.[127] The developers would qualify as consumer reporting agencies under the statute because they offered the service that gave the recommendations on the applicants' fitness for the position or provided updated information on the candidates, both of which constitute consumer reports, to help employers make hiring determinations.[128] If the algorithm was discriminatory, e.g. recommending male applicants for a position while excluding female candidates with the same or better credentials, then that is akin to the consumer reporting agency providing false information. The employers looked to the algorithm to recommend them the best candidates available but instead the

[123] Id. at 1735-36.
[124] Id. at 1736.
[125] Id.
[126] Id. at 1741.
[127] Id. at 1740.
[128] Id. at 1741.

algorithm judged applicants on the basis of a protected characteristic, eliminating qualified candidates in the process. If that occurs Thompson v. San Antonio Retail Merchants Ass'n (SARMA) shows a credit reporting agency may be liable to the candidates for negligence. In the case, the Fifth Circuit held SARMA was liable to plaintiff for negligence because SARMA, a consumer reporting agency, posted the wrong social security number for a plaintiff's profile and, as a result, the social security number's bad credit history was attributed to plaintiff.[129] Therefore, while the rejected applicant may not be able to go after the employer, they could go after the algorithm's designer as a consumer reporting agency that created inaccurate consumer reports under the Fair Credit Reporting Act.

CONCLUSION

Ending discrimination in the workplace will never be as simple as creating an algorithm and letting it run. Algorithms have their shortcomings, just as human decision-makers do. Before a designer even works on the algorithm, careful consideration needs to be given to how the target variable will be defined and measured, how training data is collected, the sources of that data, whether the data needs to be modified so it is not regurgitating prior discriminatory hiring practices, etc. Completing this step will ensure less problems crop up down the line and, if problems do occur, it takes less time to find out what went wrong.

The employer's responsibilities does not end there. Employers must periodically audit the data sets fed to the algorithms, and the audits need to be conducted by outsiders. The algorithm itself must also be continuously tested for biases. It cannot be allowed to establish a connection in its predictions between a candidate's protected characteristics such as sex and a candidate's projected competency. Features acting as proxies for protected characteristics need to be rooted out as well, if possible. If not, then steps should be taken either to hide the applicant's protected group status from the algorithm altogether or for transfer proxies to re-distribute the burden across an overlapping or larger population. As long as the employer focuses on implementing these forward-facing methods aimed at averting future discrimination and they do not set quotas for the number of protected class member hires, employers should not have to fear liability under the disparate treatment theory of employment discrimination for using these fairness measures.

[129] Id. at 1740.

That is not to say employers are completely insulated from legal liability for employment discrimination. If employers do not keep their algorithms in check, they will open themselves up to liability either under the anti-stereotyping theory of disparate treatment theory, which requires individuals not be penalized for stereotypes associated with their class, and disparate impact theory, where the algorithm was "fair in form but discriminatory in practice" and the employers refused to fix it. Companies that design the algorithm, regardless of whether they are the employers, should also have a vested interest in ensuring the algorithm remains fair, considering they constitute a consumer reporting agency under the FCRA and can be held liable to rejected applications for an algorithm's discriminatory predictions which can amount to inaccurate consumer reports.

With all of these methods to ensure algorithmic fairness at the employers' disposal and the looming threat of legal liability hanging over their heads if they ever become too lax, there is reason to believe hiring algorithms can become tools to mitigate against discrimination in hiring. As long as the balance is maintained, and humans remain engaged in the process, meaningful change can indeed be accomplished.