0001-0782/2021/January 2022
https://doi.org/10.1145/3498660

DOI:10.1145/3498660

Pauline T. Kim

▸ **James Grimmelmann,** Column Editor

Law and Technology
Addressing Algorithmic Discrimination

Considering the intersection of technical design and civil rights when building and using classification algorithms.

IT SHOULD NO longer be a surprise that algorithms can discriminate. A criminal risk-assessment algorithm is far more likely to erroneously predict a Black defendant will commit a crime in the future than a white defendant.[2] Ad-targeting algorithms promote job opportunities to race- and gender-skewed audiences, showing secretary and supermarket job ads to far more women than men.[1] A hospital's resource-allocation algorithm favored white over Black patients with the same level of medical need.[5] The list goes on. Algorithmic discrimination is particularly troubling when it affects consequential social decisions, such as who gets released from jail, or has access to a loan or health care.

Employment is a prime example. Employers are increasingly relying on algorithmic tools to recruit, screen, and select job applicants by making predictions about which candidates will be good employees. Some algorithms rely on information provided by applicants, such as résumés or responses to questionnaires. Others engage them in video games, using data about how they respond in different situations to infer personality traits. Another approach harvests information from online interactions, such as analyzing video interviews for voice patterns and facial expressions. These strategies are aimed at helping employers identify the most promising candidates, but may also reproduce or reinforce existing biases against disadvantaged groups such as women and workers of color.

Of course, human decision makers can also discriminate. Social scientists have repeatedly documented significant race- and gender-biases in human judgments. Is machine bias likely to be any worse? Perhaps not, but algorithmic selection tools raise unique concerns. Their technical nature may convey a false sense of precision, making their predictions appear objective and inevitable, rather than the product of human choices in building the model. And the discriminatory impact of algorithms is likely to scale up to a far greater extent than a biased human manager. In any case, the fact that humans are biased is no reason to ignore the problem of discriminatory algorithms.

Fortunately, the law already speaks to these concerns. For decades, it has forbidden employers from failing to hire or firing workers because of characteristics like race or sex. Although courts have not yet applied

IMAGE BY ANDRIJ BORYS ASSOCIATES, USING SHUTTERSTOCK

these laws to algorithmic selection tools, well-established legal principles offer guideposts for data scientists who build these tools and employers who deploy them.[3]

The law clearly prohibits intentional discrimination. Sometimes referred to as *disparate treatment*, this form of discrimination occurs when an employer makes an adverse decision motivated by a protected characteristic—for example, refusing to hire someone because the worker is Black or female. Given this prohibition, one might assume that avoiding discrimination is simply a matter of excluding race and other protected characteristics from algorithms' inputs and training data. Certainly, it would be unlawful to use this information *for the purpose* of building a model to exclude members of a particular group.

However, merely eliminating protected characteristics does not guarantee an algorithm will not discriminate. Under the *disparate impact* theory of discrimination, the employer can also be liable if it relies on a predictive model that has discriminatory effects. In the foundational case of *Griggs v. Duke Power Co.*[a] a group of Black workers challenged the Duke Power Company's policy that a high school diploma and minimum scores on standardized tests were required for certain skilled positions. Those requirements had the effect of barring far more Black than white workers from the higher-paying jobs because the long history of segregated schools in the region had deprived Blacks of equal educational opportunity. At the same time, the requirements bore no relationship to the skills needed to perform the jobs at issue; many white employees hired before they were implemented continued to perform satisfactorily. The Court ruled that, even if Duke Power did not intend to exclude Blacks from the skilled jobs, the company had engaged in discrimination by utilizing requirements that had an exclusionary effect and that could not be justified by legitimate business needs related to the jobs at issue.

Like the requirements in *Griggs*, a biased algorithm could produce a disparate impact even without any intent to discriminate and even when blinded to sensitive characteristics. If, for example, a training dataset has few observations of female employees, or the data include job ratings that are tainted by gender bias, then the algorithm might systematically and unjustifiably downgrade female applicants. If, as a result, it disproportionately screens out female as compared with male applicants, an employer that relies on the algorithm would face scrutiny under the disparate impact theory of discrimination.

The fact that humans are biased is no reason to ignore the problem of discriminatory algorithms.

For this reason, algorithms used for hiring and other personnel decisions should regularly be audited for their distributional effects across demographic groups. If it turns out an algorithm has a disparate impact, the employer should be able to show it accurately measures skills or attributes relevant to the particular job. In other words, the employer must demonstrate the *substantive validity* of the model. If it cannot, it should fix the tool, or stop using it.

It might seem like this situation poses an impossible dilemma: if the employer continues to use the algorithm, it will be liable for disparate impact; if it changes the model to remove a race or gender bias it may be liable for intentional discrimination. This dilemma is more apparent than real. In fact, employers are not only permitted to change practices that are found to have a disparate impact, they are required to abandon those practices if they are not justified by business necessity.

The mistaken belief that such a dilemma exists stems from a misreading of another Supreme Court case, *Ricci v. DeStefano*.[b] In that case, the New Haven

a 401 U.S. 424 (1971).

b 557 U.S. 557 (2009).

Fire Department (NHFD) discarded the results of a promotional exam after learning the racial profile of successful test takers. If it had accepted the results, the promotional class would have been almost all white, despite a diverse applicant pool, and the NHFD feared that it would be liable for disparate impact. The Court held that the NHFD discriminated against the white applicants when it decided to discard the exam results. The problem was not that the NHFD abandoned a practice with a discriminatory effect, but that it did so *after* it had publicly announced the promotional exam and many firefighters invested significant time and resources to study for and take it.

Ricci does not prohibit employers from *prospectively* changing hiring practices they discover are biased. Suppose, for example, an employer audits its hiring algorithm and learns it disproportionately screens out Black workers. *Ricci* does not prevent it from replacing that algorithm with one that is fairer for all applicants.[c] Future applicants have no fixed entitlement that an employer will always follow the same hiring process, let alone employ the same algorithm, and so no one is legally harmed if the employer adopts a different model to select candidates going forward. This is consistent with the Supreme Court's repeated emphasis on the importance of voluntary employer efforts to remove discriminatory practices, and its statements that voluntary compliance is the "preferred means" and "essential" to achieving the objectives of anti-discrimination law.[d]

c *Maraschiello v. City of Buffalo*, 709 F.3d 87 (2d Cir. 2013); *Carroll v. City of Mount Vernon*, 707 F.Supp.2d 449 (S.D. N.Y. 2010), aff'd by Carroll v. City of Mount Vernon, 453 F. App'x 99 (2d Cir. 2011).

How can an employer ensure its selection algorithm does not have unwanted discriminatory effects?

No bright line separates permissible from impermissible strategies for achieving group fairness in algorithms.

How can an employer ensure its selection algorithm does not have unwanted discriminatory effects? Computer scientists have proposed a number of approaches to mitigate bias, and these strategies generally require considering sensitive characteristics when building a model. This creates another worry for employers. What if the *debiasing effort itself* is treated as illegal discrimination because it takes sensitive characteristics like race and gender into account?

This worry is misplaced, because the law does not categorically prohibit all race- or gender-conscious actions by employers.[4] For example, in the hiring context, employers are permitted to take affirmative steps to recruit more women and racial minorities to build a broad and diverse applicant pool.[e] When designing written tests, employers have sometimes made special efforts to obtain data from underrepresented racial groups to avoid inadvertently creating biased measures, and courts have viewed these efforts favorably. Thus, while the law prohibits invidious and harmful forms of discrimination, it permits taking account of race and other sensitive characteristics in order to design selection processes that are fair to all.

This means that many strategies for debiasing algorithms are legally permissible. For example, efforts to define the target variable in a way that avoids bias should not raise any legal concerns. Similarly, designers can scrutinize the representativeness and accuracy of training data and oversample underrepresented groups or remove features that encode human biases in order to address data problems. While these strategies pay attention to race or gender effects to ensure fairness, they do not make decisions about individual applicants turn on their race or gender.

Other group fairness strategies are more vulnerable under the law. For example, courts have repeatedly criticized race and gender quotas. As a result, group fairness strategies that impose fixed proportional outcomes across subgroups—for example, by ensuring equal rates of positive outcomes for men and women regardless whether there are relevant differences between the groups—would likely trigger legal scrutiny.

Of course, gray areas remain. No bright line separates permissible from impermissible strategies for achieving group fairness in algorithms. Nevertheless, it is clear that merely blinding a selection algorithm to characteristics like race and sex is not enough to ensure nondiscrimination. And it is also clear that taking race or other protected characteristics into account in the model-building process is not categorically forbidden. While these nuances mean legal rules on discrimination are not easily reducible to code, the good news is the law currently allows considerable leeway for algorithmic designers to explore a variety of strategies for reducing or eliminating bias. C

d Ricci, 557 U.S. at 581, 583.

e For example. *Duffy v. Wolle*, 123 F.3d 1026 (8th Cir. 1997).

References

1. Ali, M. et al. Discrimination through optimization: How Facebook's ad delivery can lead to biased outcomes. In *Proceedings of the ACM on Computer-Human Interaction 3*, 199 (2019).
2. Angwin, J. et al. Machine bias. ProPublica (2016); https://bit.ly/3c1zXE2
3. Kim, P.T. Data-driven discrimination at work. *William & Mary Law Review 58* (2017), 857.
4. Kim, P.T. Race-aware algorithms: Fairness, nondiscrimination and affirmative action. *California Law Review 110* (2022).
5. Obermeyer, Z. et al. Dissecting racial bias in an algorithm used to manage the health of populations. *Science 366* (2019), 447.

Pauline T. Kim (kim@wustl.edu) is the Daniel Noyes Kirby Professor of Law, Washington University School of Law, St. Louis, MO, USA.

This work was supported by NSF-Amazon award # IIS-1939677.