AEA Papers and Proceedings 2018, 108: 22–27
https://doi.org/10.1257/pandp.20181018

ADVANCES IN BIG DATA RESEARCH IN ECONOMICS

Algorithmic Fairness[†]

By Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan*

The growing use of algorithms in social and economic life has raised a concern: that they may inadvertently discriminate against certain groups. For example, one recent study found that natural language processing algorithms can embody basic gender biases, such as associating the word *nurse* more closely with the word *she* than with the word *he* (Caliskan, Bryson, and Narayanan 2017). Because the data used to train these algorithms are themselves tinged with stereotypes and past discrimination, it is natural to worry that biases are being "baked in."

We consider this problem in the context of a specific but important case, one that is particularly amenable to economic analysis: using algorithmic predictions to guide decisions (Kleinberg et al. 2015). For example, predictions about a defendant's safety risk or flight risk are increasingly being proposed as a means to guide judge decisions about whether to grant bail. Discriminatory predictions in these cases could have large consequences. One can easily imagine how this could happen since recidivism predictions will be polluted by the fact that past arrests themselves may be racially biased. In fact, a recent *ProPublica* investigation argued that the risk tool used in one Florida county was in fact discriminatory (Angwin et al. 2016). This widely-read article helped further elevate concerns about fairness within the policy and research communities alike, with subsequent work showing that the trade-offs are more subtle than was initially apparent.[1]

These concerns have led to a large literature that tries to "blind" the algorithm to race to avoid exacerbating existing unfairnesses in society. Numerous studies (many of them in computer science) have pointed out that this requires more than just excluding race from the predictor, since protected features such as race could be reconstructed from other features. To solve this "reconstruction problem," procedures have been proposed such as pre-processing the data to orthogonalize the explanatory variables ("inputs") or outcomes to race, or modifying the loss function the algorithm seeks to optimize to penalize race disparities in outcomes.

We argue that this perspective about how to promote algorithmic fairness, while intuitive, is misleading and in fact may do more harm than good. We develop a simple conceptual framework that models how a social planner who cares about equity should form predictions from data that may have potential racial biases. Our primary result is exceedingly simple, yet often overlooked: a preference for fairness should not change the choice of estimator. Equity

* Kleinberg: Cornell University, Ithaca, NY 14853 (email: kleinber@cs.cornell.edu); Ludwig: University of Chicago, 1155 East 60th Street, Chicago, IL 60637, and NBER (email: jludwig@uchicago.edu); Mullainathan: Harvard University, 1805 Cambridge Street, Cambridge, MA 02138, and NBER (email: mullain@fas.harvard.edu); Rambachan: Harvard University, 1805 Cambridge Street, Cambridge, MA 02138 (email: asheshr@g.harvard.edu). Thanks to the NSF for support to Rambachan, to the Simons Foundation for support of Kleinberg, and to the Arnold Foundation, Susan and Tom Dunn, and Ira Handler for providing support to the University of Chicago Crime Lab and Urban Labs. Any errors and all opinions are of course our own.

[†] Go to https://doi.org/10.1257/pandp.20181018 to visit the article page for additional materials and author disclosure statement(s).

[1] Angwin et al. (2016) argued the risk tool they examined was biased because African Americans are more likely to be mis-classified as higher risk, while whites were more likely to be mis-classified as lower risk. Kleinberg, Mullainathan, and Raghavan (2017) and Chouldechova (2017) note that this finding is an unavoidable consequence for calibrated risk tools in the presence of differences in offending rates across groups, unless we have tools that are perfectly predictive of risk. An alternative measure of fairness used in Kleinberg et al. (2018) focuses on a quantity derived from the actual decision outcome: detention rates to African Americans.

preferences can change how the estimated prediction function is used (such as setting a different threshold for different groups) but the estimated prediction function itself should not change. Absent legal constraints, one should include variables such as gender and race for fairness reasons. As we show in an empirical example below, the inclusion of such variables can increase both equity and efficiency.

Our argument collects together and builds on existing insights to contribute to how we should think about algorithmic fairness.[2] This argument is not specific to machine learning—indeed the argument is cleanest, and presented here, within the context of unbiased estimators. Additional issues arise beyond those we consider here for high-dimensional estimation procedures that trade off bias and variance to maximize out-of-sample (OOS) prediction accuracy, which we discuss separately in other work.

We empirically illustrate this point for the case of using predictions of college success to make admissions decisions. Using nationally representative data on college students, we underline how the inclusion of a protected variable—race in our application—not only improves predicted GPAs of admitted students (efficiency), but also can improve outcomes such as the fraction of admitted students who are black (equity). The reason for this result is extremely simple. Equity preferences involve increasing the fraction of black applicants admitted. Within that set, society is still served best by ranking as well as possible using the best possible predictions. Forming the best predictions possible aids both equity and efficiency.

I. Conceptual Framework

As an illustrative case of our framework, suppose we are interested in a social planner that is trying to make college admissions decisions based on anticipated college success. To do so, she applies some procedure to historical data to form a predictor of college success and uses that predictor to decide on future admissions. Individuals are described by (Y, X, R), where Y is eventual (measured) college success, X is a set of academic variables that are observed at time of admissions, and $R \in \{0, 1\}$ is the applicant's race (with $R = 1$ if the individual is from the minority group of interest).

We consider an *efficient planner* and an *equitable planner*. The efficient planner maximizes an objective function $\phi(S)$ that depends only on the predicted performance of the set S of admitted students.

We assume that the efficient planner applies an estimator to a given dataset on individuals consisting of (Y_i, X_i, R_i) to produce a predictor $\hat{f}(X, R)$. We say this objective function ϕ is *compatible* with the prediction function $\hat{f}$ if the following natural monotonicity condition holds: If S and S' are two sets of students of the same size, sorted in descending order of predicted performance $\hat{f}(X, R)$, and the predicted performance $\hat{f}(X, R)$ of the ith student in S is at least as large as the predicted performance of the ith student in S' for all i, then $\phi(S) \geq \phi(S')$.[3] Given an objective function and a compatible predictor, the efficient planner has a simple rule. For a desired number of admitted students K, the efficient planner simply admits the set S consisting of the K students with the highest $\hat{f}(X, R)$ values.

Now consider an equitable planner, who has preferences over both grades and the racial composition of the admitted class. They seek to maximize $\phi(S) + \gamma(S)$, where ϕ is compatible with $\hat{f}$ as before, and $\gamma(S)$ is monotonically increasing in the number of students in S who have $R = 1$ (and thus belong to the minority group).[4]

How should the equitable planner solve her optimization problem? The following theorem

[2] See, for example, the excellent discussions of existing algorithmic fairness research in Barocas and Selbst (2016) and Dwork et al. (2012, 2017), whose arguments are consistent with the view we take in the present work. The importance of within-group rankings for affirmative action has been noted by Fryer and Loury (2013). Similarly, Corbett-Davies et al. (2017) show that several notions of algorithmic fairness in the context of criminal justice rely on the use of race-specific decision rules.

[3] A familiar case of compatibility is when ϕ is simply the sum of performance of admitted students and OLS is applied to the data to produce an unbiased estimator, such that for an individual i, $Y_i = \hat{f}(X_i, R_i) + \epsilon_i$, where ϵ_i is mean zero and orthogonal to (X_i, R_i).

[4] The analysis that follows extends directly to the more general case in which the equitable planner seeks to maximize $\psi(S) + \gamma(S)$, where ψ is compatible with f but may be different from the efficient planner's function ϕ. We use the case in which the efficient and equitable planners have the same ϕ in our exposition.

shows that they should also rank by $\hat{f}$, just as the efficient planner, but change the cutoffs used for admission for the two groups defined by $R = 0$ and $R = 1$.

THEOREM 1: *For some choice of K_0 and K_1 with $K_0 + K_1 = K$, the equitable planner's problem can be optimized by choosing the K_0 applicants in the $R = 0$ group with the highest $\hat{f}(X,R)$, and the K_1 applicants in the $R = 1$ group with the highest $\hat{f}(X,R)$.*

We can sketch the proof of this theorem as follows. Let S^* be any set of K applicants that maximizes $\phi(S) + \gamma(S)$, and partition S^* into the applicants S_0^* in the $R = 0$ group and S_1^* in the $R = 1$ group. Let $K_0 = |S_0^*|$ and $K_1 = |S_1^*|$; let S_0^+ be the K_0 applicants in the $R = 0$ group with the highest $\hat{f}(X,R)$, and let S_1^+ be the K_1 applicants in the $R = 1$ group with the highest $\hat{f}(X,R)$. Write $S^+ = S_0^+ \cup S_1^+$. One can prove that if we sort S^+ and S^* in descending order of predicted performance, the ith student in S^+ has predicted performance at least as large as the ith student in S^*. Hence by the compatibility of ϕ and f, we have $\phi(S^+) \geq \phi(S^*)$. Since S^+ and S^* have the same number of members with $R = 1$ by construction, we also have $\gamma(S^+) = \gamma(S^*)$; thus $\phi(S^+) + \gamma(S^+) \geq \phi(S^*) + \gamma(S^*)$, and so S^+ is a set maximizing the equitable planner's objective function and satisfying the conditions of the theorem.

Put in another way, given the equitable planner's preferences, they still wish to rank-order individuals within each group using the same estimate of expected performance $\hat{f}(X,R)$.

The intuition behind this Theorem is simple. The efficient planner only values ranking on the best possible prediction of output. An equitable planner, conditional on the fraction of minority students admitted, cares about this as well. Since the fraction of admitted students that are minorities can always be altered by changing the thresholds used for admission, the equitable planner should use the same prediction function as the efficient planner. Implicit in this theorem is that the use of race will always be strictly improving for the equitable planner's objective function as long as race is useful for predicting Y. This happens exactly when we feel there is disadvantage—when individuals of $R = 1$ have a different process than those with $R = 0$. In this case, access to R improves prediction quality.

II. Further Implications

The theorem in the previous section helps parse some common reasons we worry that the data may bake in bias. First, we may worry that the inputs (X) are biased. In the college context, America's history of segregated schools may affect the degree to which minority students are comparably prepared to succeed in college. One consequence could be that with fewer inputs, black students are less prepared for college, so that $E[Y|R = 1] < E[Y|R = 0]$. Implicit in our theorem is that the solution here is to set a different threshold for admissions. Another concern may be that, even for the same level of preparation to succeed in college, black applicants appear worse on observed inputs. For example, they may receive less coaching on how to take standardized tests like the SAT. Yet this scenario implies that $f(X,R)$ crucially depends on R. For the algorithm to account for this sort of bias, it *must* know R. If white students are given more SAT prep, then the same SAT score implies *higher* college success for a black student than a white one; that is, $E[Y|X, R = 1] > E[Y|X, R = 0]$. A prediction function can only uncover this if it is allowed to put different "slopes" on the SAT score for white and black candidates.[5] As a result, racial equity is promoted by giving the algorithm access to race.

Second, we may also worry that the Y variables themselves are biased. When we measure Y using college GPA, biases faced in the college experience are reproduced in Y. Specifically, there may be a true Y^* (such as actual learning in college) which determines ϕ, even though we only measure Y (grades). We may fear that $E[Y - Y^*|R = 1] < E[Y - Y^*|R = 0]$. Even here, as long as $\hat{f}$ is compatible to ϕ in the sense above, the theorem says the equitable and efficient planners should use the same $\hat{f}$. The

[5] This discussion subsumes one where there are true X^*, and the X are racially biased measures of X^*. This type of mis-measurement will have the same effect.

intuition is again straightforward. As long as Y and Y^* are monotonically related, ranking on $\hat{f}$ remains the best strategy even when you care about equity. The $Y - Y^*$ bias can be accounted for by setting a different threshold for the discriminated group.

III. Data

We rely on the public-use version of the US Department of Education's National Education Longitudinal Study of 1988 (NELS:88). This dataset captured information for a nationally representative sample of students who entered eighth grade in the fall of 1988, and who were then followed up in 1990, 1992, 1994, and finally in 2000 (when respondents were in their mid-20s).[6]

The decision we examine is college admission, specifically whether to admit a student to a four-year college or university. We limit our analysis sample to those who were followed through the 2000 survey wave and had ever attended a four-year institution. To simplify, we focus just on two groups: non-Hispanic white students ($N = 4{,}274$) and black students ($N = 469$).

We assume the admit decision is based on predicted student performance (college grade point average). Predictors taken from the 1988, 1990, and 1992 waves of NELS data include (besides race) high school grades, course taking patterns, extracurricular activities, and student performance on the standardized achievement tests that NELS administered to students in four core subject areas: math, reading, science, and social studies.[7] Consistent with previous studies, college graduation rates are higher for white students than black students in our sample (67.4 percent versus 50.9 percent). College grades such as share earning a GPA of at least 2.75 are also higher on average for white than black students, 82.2 percent versus 69.5 percent, as are most of our measures of high school outcomes.[8]

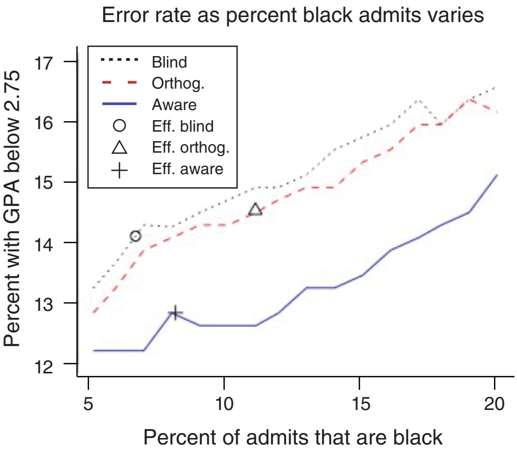

Figure 1. Racial Composition/Grades Curves

IV. Empirical Results

We summarize our main results by showing what would happen to college admissions outcomes for both the efficient planner and equitable planner using different candidate prediction functions. The binary outcome we predict has $Y = 1$ if the student's college GPA is <2.75. We consider predictions from a simple unbiased algorithm (ordinary least squares), one version of which is blinded to race altogether, one of which pre-processes the inputs to make them orthogonal to race, and one of which is made race-aware by interacting race with the various predictors described above.[9]

For starters, Figure 1 shows the efficient planner would choose to use the race-aware predictor. The cross in the figure shows what would happen if the efficient planner selected the top 50 percent of four-year college students in the NELS using the race-aware predictor (that is, by rank-ordering all four-year students in the NELS by the predicted outcome from the race-aware predictor, then selecting the top half). Just under

[6] The NELS provides sampling weights to account for the fact that not all baseline respondents were eligible for follow-up waves. We present unweighted results below, but our findings are not sensitive to using the weights.

[7] The public-use NELS tells us whether a student took the SAT or ACT, but not their score, so we use these tests as a proxy.

[8] We do not use data on the sociodemographics of the student's family or school in any analyses.

[9] Results are qualitatively similar when we instead use a machine learning algorithm (random forest), or use different outcome measures for college performance such as varying the GPA threshold or an indicator for college completion.

13 percent of students admitted using the race-aware algorithm would go on to get GPAs below 2.75, at least a full percentage point lower than if the efficient planner had instead rank-ordered students for admission using either the race-blind algorithm or the predictor that first orthogonalizes inputs to race (the circle and triangle in Figure 1, respectively). Because the efficient planner cares only about efficiency (i.e., location along the y-axis), using the race-aware predictor dominates.

More interesting is our evidence that the equitable planner (who uses a different threshold to admit white versus black students, to promote fairness) would also wish to use the race-aware predictor. Varying the threshold used for black students changes the share of the admitted 50 percent of students who are black (shown on the x-axis): for a given predictor the lower the threshold used for black students, the larger the share of admitted students who are black, but the higher the share of admitted students who achieve a GPA <2.75. So, for a given predictor, the curve shows that the possible combinations of diversity and college achievement that can be achieved has a positive slope. The race-aware predictor dominates the other prediction functions, even for the equitable planner. For any given level of diversity among admitted students, using the race-aware predictor leads to the smallest share of admitted students with low grades. Equivalently, for any given level of achievement among admitted students, using the race-aware predictor would lead to admission of relatively more black applicants.

Figure 2 shows why this is so. The race-blind predictor mis-ranks black students. This “heatmap” shows the distribution of black students in our NELS sample across predicted-outcome deciles according to the race-blind (x-axis) or race-aware (y-axis) predictors. For example, consider the group of students in the southeast part of Figure 2. The race-blind predictor classifies them as being in the ninth decile of predicted probability of receiving a GPA <2.75, but, according to the race-aware predictor, they are actually members of the lowest decile.

This mis-ranking stems from the differences in the slopes for the relationship between college grades and different predictors (such as high school achievement tests) for black versus white students. Only by giving the algorithm access to race can we account for this, improve the rank-ordering within the pool of black applicants and form a decision rule that dominates those based on race-blind predictors.

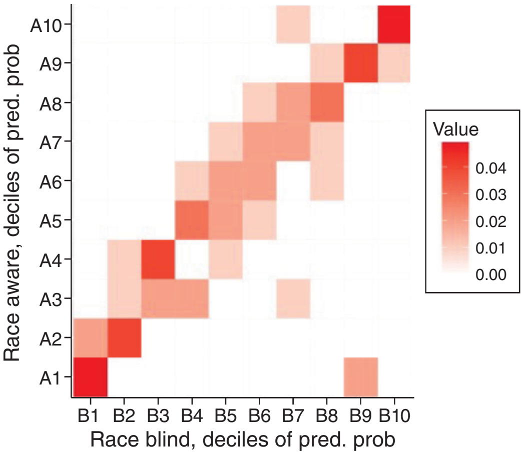

Figure 2. Heatmap of Rankings of Black Applicants by Predicted Probability of GPA < 2.75, Using Race-Aware Versus Race-Blind Algorithms

V. Conclusion

Concern about the potential fairness consequences of algorithmic decision-aids is understandable and plays an important role in debates about their wide-scale adoption. Our central argument is that across a wide range of estimation approaches, objective functions, and definitions of fairness, the strategy of blinding the algorithm to race inadvertently detracts from fairness.

REFERENCES

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

Barocas, Solon, and Andrew D. Selbst. 2016. “Big Data’s Disparate Impact.” *California Law Review* 104: 671–732.

Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan. 2017. “Semantics Derived Automatically from Language Corpora Contain

Human-Like Biases." *Science* 356 (6334): 183–86.

Chouldechova, Alexandra. 2017. "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments." *Big Data* 5 (2): 153–63.

Corbett-Davies, Sam, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. "Algorithmic Decision Making and the Cost of Fairness." *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*: 797–806.

Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. "Fairness through Awareness." *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*: 214–26.

Dwork, Cynthia, Nicole Immorlica, Adam Tauman Kalai, and Max Leiserson. 2017. "Decoupled Classifiers for Fair and Efficient Machine Learning. *arXiv* 1707.06613.

Fryer, Roland G., Jr., and Glenn C. Loury. 2013. "Valuing Diversity." *Journal of Political Economy* 121 (4): 747–74.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. "Prediction Policy Problems." *American Economic Review* 105 (5): 491–95.

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. 2017. "Inherent Trade-Offs in the Fair Determination of Risk Scores." *Proceedings of the 8th Conference on Innovation in Theoretical Computer Science*: 43:1–43:23.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. "Human Decisions and Machine Predictions." *Quarterly Journal of Economics* 133 (1): 237–93.