AEA Papers and Proceedings 2020, 110: 91–95
https://doi.org/10.1257/pandp.20201036

An Economic Perspective on Algorithmic Fairness[†]

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

There are widespread concerns that the growing use of machine learning algorithms in important decisions may reproduce and reinforce existing discrimination against legally protected groups. Most of the attention to date on issues of "algorithmic bias" or "algorithmic fairness" has come from computer scientists and machine learning researchers, whose fields typically focus on the design of machine learning algorithms.[1] Perhaps as a result, the focus of this literature has largely been on how concerns about algorithmic fairness relate to the design of machine learning algorithms themselves.

In this paper, we argue that concerns about algorithmic fairness are at least as much about questions of how discrimination manifests itself in data, decision-making under uncertainty, and optimal regulation. For example, how do biases in data propagate into predictions? Do biased data necessarily lead to biased algorithms? Should algorithms have access to protected group characteristics? To fully answer questions like these, an economic framework is necessary—and as a result, economists have much to contribute.

A Class of Prediction Policy Problems.—Throughout this paper, we focus on a particular class of prediction policy problems that we refer to as "screening decisions." In a screening decision, a decision-maker must select one or more people from a larger pool on the basis of a prediction of an unknown outcome of interest, Y^*. These types of decisions are widespread and of enormous economic importance. For example, pretrial release decisions depend on the probability that the defendant will fail to appear in court or recidivate (Kleinberg, Lakkaraju, et al. 2018). Firms base hiring decisions upon predictions of the future productivity of applicants (Chalfin et al. 2016). Further examples include credit approvals (Fuster et al. 2018), medical testing (Mullainathan and Obermeyer 2019), and college admissions (Kleinberg, Ludwig, et al. 2018).[2] Screening decisions are an ideal application of supervised machine learning algorithms because data on past people and decisions can be used to build algorithms that generate predictions of the outcome of interest as a function of observable features of people, W.

I. How Do Algorithms Interact with Existing Disparities and Discrimination?

A. *Decomposing Average Group Differences in Predictions*

Suppose that a firm has a gender disparity in hiring because an algorithm produces different predictions for men versus women. Typically the default assumption of the existing literature in computer science is that the differences in the distribution of predictions by group must stem from some problem of bias with the algorithm itself.

In contrast, economists, when faced with such a disparity in hiring outcomes, often begin with an accounting exercise that decomposes the observed disparity in hiring rates into different

* Rambachan: Harvard University (email: asheshr@g.harvard.edu); Kleinberg: Cornell University (email: kleinber@cs.cornell.edu); Ludwig: University of Chicago and NBER (email: jludwig@uchicago.edu); Mullainathan: University of Chicago and NBER (email: Sendhil.Mullainathan@chicagobooth.edu). Rambachan gratefully acknowledges financial support from the National Science Foundation Graduate Research Fellowship (grant DGE1745303).

[†] Go to https://doi.org/10.1257/pandp.20201036 to visit the article page for additional materials and author disclosure statement(s).

[1] Barocas, Hardt, and Narayanan (2019) is a publicly available textbook introduction to this computer science literature, and Cowgill and Tucker (2019) provides a survey for economists.

[2] Our focus on screening decisions excludes some applications of unsupervised learning algorithms in which concerns of algorithmic bias arise (Caliskan, Bryson, and Narayanan 2017).

features of the observed data.[3] Similarly, we can decompose the algorithm's observed behavior into several components. Let $\tilde{Y}$ denote the measured outcome predicted by the algorithm, where

$$(1) \qquad \tilde{Y} = \underbrace{Y^*}_{\text{outcome of interest}} + \underbrace{\Delta_Y}_{\text{measurement error}}.$$

In general, the measured outcome may not equal the outcome of interest Y^*. The raw difference in the algorithm's estimated predictions, $\hat{E}\big[\tilde{Y} \,|\, G = 1\big] - \hat{E}\big[\tilde{Y} \,|\, G = 0\big]$, can be written as

$$\underbrace{\big(E[Y^*|G = 1] - E[Y^*|G = 0]\big)}_{\text{base rate differences}} + \underbrace{\big(E[\Delta_Y|G = 1] - E[\Delta_Y|G = 0]\big)}_{\text{measurement error differences}} + \underbrace{\big(\hat{\varepsilon}(1) - \hat{\varepsilon}(0)\big)}_{\text{estimation error differences}},$$

where $\hat{\varepsilon}(g) = \hat{E}\big[\tilde{Y} \,|\, G = g\big] - E\big[\tilde{Y} \,|\, G = g\big]$ is the algorithm's estimation error on group g. In words, a raw difference in predictions across groups may arise for three reasons: differences in base rates, differences in measurement error, or differences in estimation error across groups. Intervening at the level of the algorithm may address only the last two components of the disparity by investing in better training data and collecting a better proxy for the outcome of interest. In contrast, the base rate difference is a product of the underlying socioeconomic context itself, not the algorithm. Addressing this component may require more fundamental policy interventions that seek to make changes further upstream to address the differences in observable characteristics between men and women. Put differently, such decompositions can help better target policy interventions.

B. *Can Algorithms Arbitrage Discrimination?*

In existing work on algorithmic bias, there is a common intuition that algorithms necessarily replicate the bias of the socioeconomic environment. This intuition is typically summarized with the phrase "bias in, bias out."

As a simple example, imagine judges are discriminatory against African American defendants in pretrial release decisions—judges are less likely to release an African American defendant than a white defendant who has a similar probability of failing to appear in court (Arnold, Dobbie, and Yang 2018). Would such an algorithm trained to predict failure-to-appear rates on past releases reproduce the judges' bias? No. Because discriminatory judges apply a "higher threshold" for releasing African American defendants, released African American defendants are actually *lower* risk than released white defendants, and an algorithm would learn this from past releases.

Rambachan and Roth (2019) provides a strong comparative static, which shows that this simple intuition holds under quite general conditions. If training data are created only if discriminatory human decision-makers take some action[4] (e.g., we observe whether defendants fail to appear in court only if a judge releases them), and the human decision-maker has access to unobservables, then algorithmic decision-making may *reverse* bias. The more discriminatory the human decision-maker is against a protected group, the more favorable the resulting algorithm is toward that group.

This result mirrors an old insight in the economics of discrimination that equilibrium forces may work to reduce existing discrimination. For example, in labor markets, discriminating firms may be competed out of existence as profit-maximizing firms expand and arbitrage away wage differences across groups (Becker 1957).[5] As seen, a similar principle may apply in algorithmic decision-making, where algorithms play an analogous role to profit-maximizing firms through their data-driven optimization.

The work ahead is to better understand how these two forces—whether algorithms magnify bias or reduce bias—trade off against each other in richer environments that incorporate richer models of human biases and dynamics.

[3] Fortin, Lemieux, and Firpo (2011) reviews such decomposition methods in economics.

[4] This is commonly referred to as the "selective labels problem" in the machine learning literature (Kleinberg, Lakkaraju, et al. 2018).

[5] Other theories suggest alternative mechanisms through which discrimination may persist (see Rodgers 2006).

II. Do Equity Preferences Modify the Design of Algorithms?

So far, we have focused on how the underlying data-generating process affects predictions, possibly resulting in “biased” algorithms. How then should an equity-minded social planner go about building an algorithm from some observed training data?

This problem is often the starting point of existing research on algorithmic fairness. Existing work assumes that the social planner wishes to construct the most accurate prediction function among those that are defined to be “fair,” where fairness is formalized as an additional constraint that prediction functions must satisfy.[6]

In contrast, an economic approach models fairness as a property of the social planner’s *preferences*, not as a further constraint in her optimization problem. Kleinberg, Ludwig, et al. (2018) and Rambachan et al. (2020) define a social welfare function that depends on the outcomes produced by the screening decision. The social welfare function captures an explicit preference for more equitable outcomes across groups.

Starting from this perspective, these papers provide two versions of an *equity irrelevance result*—equity preferences affect only the screening rule, not the prediction function itself.[7] A prediction function simply aggregates information, summarizing the observed relationship between the outcome of interest and the observed features. Modifying the prediction function by removing features, blinding it to protected group characteristics, or introducing additional constraints in the training procedure may throw away potentially useful information.

III. Defining the Objective Function

While the equity irrelevance results offer strong null results, they are limited by their assumptions, and it is not obvious that they apply in all applications of algorithmic decision-making. For example, in many screening decisions, there is no precise definition of the outcome of interest. In existing theoretical work in computer science, exactly how the measured outcome is defined is often left unmodeled. In practice, the choice of the outcome to be predicted is made for convenience and is left to the data scientist. Yet this choice can have large effects on the resulting algorithm and its properties.

Consider Obermeyer et al. (2019), which investigates an algorithm that affects the care of millions of patients across the United States. It aims at directing the sickest patients into care coordination programs. However, there are large racial disparities in enrollment decisions based upon the algorithm’s predictions. Such disparities arise because the algorithm predicts *observed costs*, not a measure of health. While observed costs may appear to be a reasonable proxy for patient health, as sicker patients are also likely to be more costly patients, it was precisely this choice that created large racial inequities. One interpretation of this example is that health is a latent variable, which manifests itself in high-dimensional data, and so there is no obvious scalar proxy. Observed cost was chosen simply for convenience.

More generally, Mullainathan and Obermeyer (2017) provides a framework for understanding the implications of mismeasured proxies for the evaluation of machine learning algorithms. A core challenge is that one prediction function may appear to dominate another because it accurately predicts the proxy’s idiosyncratic error, not the outcome of interest. Much work remains to be done on developing robust training procedures in the presence of mismeasured proxies.

IV. Toward a Theory of Optimal Algorithmic Regulation

An alternative interpretation of the example in Obermeyer et al. (2019) is that cost is actually the correct outcome of interest for hospitals and health insurance providers, but it is not the correct outcome of interest for society. The

[6] Canonical papers include Dwork et al. (2012); Zemel et al. (2013); and Hardt, Price, and Srebro (2016).

[7] Corbett-Davies et al. (2017); Lipton, McAuley, and Chouldechova (2018); and Menon and Williamson (2018) provide analogs of this result in the computer science literature by deriving the solution to a fairness-constrained loss-minimization problem. In contrast, these equity irrelevance results characterize the unconstrained optimum for a wide class of social welfare functions.

objective of hospitals and health insurance providers, which designed and implemented the predictive algorithm, is to maximize profits, but the social welfare function is defined over patient health.

This competing explanation points to a larger, understudied problem in algorithmic fairness: developing a theory of optimal algorithmic regulation. In many applications, the social planner interacts with third-party firms that control the construction of the prediction function and the screening rule. These firms do not share the same preferences as the social planner, and some may even wish to discriminate against protected groups. In this sense, many applications of algorithmic decision-making are better modeled as a *regulation problem* in which the social planner's inability to directly dictate choices leaves us in a second-best world.

Rambachan et al. (2020) provides a simple model of such a regulation problem in which an equity-minded social planner faces a market of firms that face their own screening decision. The social planner may affect the screening decisions of the firms only through "model regulations," meaning that she can ban certain characteristics from being used in screening decisions. The authors show that the effects of algorithmic decision-making on this regulation problem depend on what firms must disclose to the social planner about their algorithms. If the full underlying training data and training procedure must be disclosed in what the authors call an "algorithmic audit," then it is optimal to let firms use any characteristic that is predictive of the outcome of interest. This is reminiscent of the earlier equity irrelevance results.

Studying the design of optimal policy is a core question throughout economics, from the design of tax systems to the regulation of monopolies. Moving forward, tools from mechanism design will be useful in characterizing properties of optimal algorithmic regulation in full generality.

V. Conclusion

Fears about algorithmic bias are widespread. While at first glance the core questions surrounding algorithmic bias appear to be questions for computer scientists, we argue that economists are particularly well equipped to play an important role in this key policy domain moving forward.

REFERENCES

Arnold, David, Will Dobbie, and Crystal S. Yang. 2018. "Racial Bias in Bail Decisions." *Quarterly Journal of Economics* 133 (4): 1885–932.

Barocas, Solon, Moritz Hardt, and Arvind Narayanan. 2019. "Fairness and Machine Learning: Limitations and Opportunities." https://fairmlbook.org/ (accessed December 20, 2019).

Becker, Gary S. 1957. *The Economics of Discrimination*. Chicago: University of Chicago Press.

Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan. 2017. "Semantics Derived Automatically from Language Corpora Contain Human-Like Biases." *Science* 356 (6334): 183–86.

Chalfin, Aaron, Oren Danieli, Andrew Hillis, Zubin Jelveh, Michael Luca, Jens Ludwig, and Sendhil Mullainathan. 2016. "Productivity and Selection of Human Capital with Machine Learning." *American Economic Review* 106 (5): 124–27.

Corbett-Davies, Sam, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. "Algorithmic Decision Making and the Cost of Fairness." *KDD '17: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 23: 797–806.

Cowgill, Bo, and Catherine Tucker. 2019. "Economics, Fairness and Algorithmic Bias." http://dx.doi.org/10.2139/ssrn.3361280.

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

Fortin, Nicole, Thomas Lemieux, and Sergio Firpo. 2011. "Decomposition Methods in Economics." In *Handbook of Labor Economics*, Vol. 4A, edited by Orley Ashenfelter and David Card, 1–102. Amsterdam: North-Holland.

Fuster, Andreas, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther. 2018. "Predictably Unequal? The Effects of Machine Learning on Credit Markets." https://dx.doi.org/10.2139/ssrn.3072038.

Hardt, Moritz, Eric Price, and Nathan Srebro. 2016. "Equality of Opportunity in Supervised Learning." *NIPS 2016: Proceedings of the 30th International Conference on Neural Information Processing Systems* 30: 3323–31.

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

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. 2018b. "Algorithmic Fairness." *AEA Papers and Proceedings* 108: 22–27.

Lipton, Zachary, Julian McAuley, and Alexandra Chouldechova. 2018. "Does Mitigating ML's Impact Disparity Require Treatment Disparity?" *NIPS 2018: Proceedings of the 32nd Conference on Neural Information Processing Systems* 32: 2–11.

Menon, Aditya Krishna, and Robert C. Williamson. 2018. "The Cost of Fairness in Binary Classification." *Proceedings of Machine Learning Research* 81: 107–18.

Mullainathan, Sendhil, and Ziad Obermeyer. 2017. "Does Machine Learning Automate Moral Hazard and Error?" *American Economic Review* 107 (5): 476–80.

Mullainathan, Sendhil, and Ziad Obermeyer. 2019. "A Machine Learning Approach to Low-Value Health Care: Wasted Tests, Missed Heart Attacks and Mis-predictions." NBER Working Paper 26168.

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations." *Science* 366 (6464): 447–53.

Rambachan, Ashesh, Jon Kleinberg, Jens Ludwig, and Sendhil Mullainathan. 2020. "An Economic Approach to Regulating Algorithms." https://scholar.harvard.edu/asheshr/publications/economic-approach-regulating-algorithms.

Rambachan, Ashesh, and Jonathan Roth. 2019. "Bias In, Bias Out? Evaluating the Folk Wisdom." https://arxiv.org/abs/1909.08518.

Rodgers, William M. III, ed. 2006. *Handbook on the Economics of Discrimination*. Cheltenham, UK: Edward Elgar.

Zemel, Rich, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. "Learning Fair Representations." *Proceedings of the 30th International Conference on Machine Learning* 28 (3): 325–33.