Fair Prediction with Endogenous Behavior

CHRISTOPHER JUNG, University of Pennsylvania, USA
SAMPATH KANNAN, University of Pennsylvania, USA
CHANGHWA LEE, University of Pennsylvania, USA
MALLESH PAI, Rice University, USA
AARON ROTH, University of Pennsylvania, USA
RAKESH VOHRA, University of Pennsylvania, USA

There is great interest in whether machine learning algorithms deployed in consequential domains (e.g. in criminal justice) treat different demographic groups "fairly." However, there are several proposed notions of fairness, typically mutually incompatible. Using criminal justice as an example, we study a model in which society chooses an incarceration rule. Agents of different demographic groups differ in their outside options (e.g. opportunity for legal employment) and decide whether to commit crimes. We show that equalizing type I and type II errors across groups is consistent with the goal of minimizing the overall crime rate; other popular notions of fairness are not.

CCS Concepts: • **Theory of computation → Algorithmic game theory**; • **Applied computing → Economics**.

Additional Key Words and Phrases: machine learning; fairness; calibration; risk score; criminal justice

ACM Reference Format:
Christopher Jung, Sampath Kannan, Changhwa Lee, Mallesh Pai, Aaron Roth, and Rakesh Vohra. 2020. Fair Prediction with Endogenous Behavior. In *Proceedings of the 21st ACM Conference on Economics and Computation (EC '20), July 13–17, 2020, Virtual Event, Hungary.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3391403.3399473

Algorithms to automate consequential decisions are frequently suspected of being unfair or discriminatory. The suspicions are not hypothetical. Discoveries of "algorithmic bias" have prompted policymakers, regulators, and computer scientists to propose that algorithms be designed to satisfy notions of fairness.

What measure(s) of fairness should designers be held to, and how do these constraints interact with the original objectives the algorithm was designed to target? This question has inspired a literature proposing (or criticizing) notions of fairness based on ethical/normative grounds. The literature evaluates algorithms on the basis of these measures and/or proposes novel algorithms that better trade-off the goals of the original designer with these fairness desiderata. However, various fairness measures are often incompatible with one another. For example, enforcing parity of false positive or false negative rates for e.g. parole decisions typically requires making parole decisions using different thresholds on the posterior probability that an individual will commit a crime for different groups which can be seen as unfair. Furthermore, these notions of fairness are sometimes disconnected from and lead to unpalatable tradeoffs with other economic and social objectives one might care about. Lastly, the literature frequently assumes that the agent types,

EC '20, July 13–17, 2020, Virtual Event, Hungary

ACM ISBN 978-1-4503-7975-5/20/07.
https://doi.org/10.1145/3391403.3399473

which are relevant to the decision at hand, are exogenously determined, i.e. unaffected by the decision rule that is selected. When agent decisions are exogenously fixed, it has been observed that optimizing natural notions of welfare and accuracy (incarcerating the guilty, acquitting the innocent) are achieved by decision rules that select a uniform threshold on "risk scores" that are well calibrated — for example, the posterior probability of criminal activity — which tend *not* to satisfy statistical notions of fairness that have been proposed in the literature.

Here we consider a setting in which agent decisions are endogenously determined and show that parity of false positive and negative rates (known as *equalized odds*) *is* aligned with the natural objective of minimizing crime rates, while parity of positive predictive value and posterior threshold uniformity are *not*. We focus on the objective of minimizing crime rates as opposed to maximizing accuracy, which is what the previous literature has mostly focused on. Although the model need not be tied to any particular application, we develop it using the language of criminal justice. We treat agents as rational actors whose decisions about whether or not to commit crimes are *endogenously* determined as a function of the incentives given by the decision procedure society uses to punish crime. The possibility for unfairness arises because agents are ex-ante heterogeneous: their demographic group is correlated with their underlying incentives— for example each individual has a private *outside option* value for not committing a crime, and the distribution of outside options differs across groups. Our key result is that policies that are optimized to minimize crime rates are compatible with a popular measure of demographic fairness — equalizing false positive and negative rates across demographics — and are generally incompatible with equalizing positive predictive value and uniform posterior thresholds. Thus, which of these notions of fairness is compatible with natural objectives hinges crucially on whether one believes that criminal behavior is responsive to policy decisions or not.

Our results have direct implications for regulatory testing for unfairness. A regulator who wishes to test whether an adjudicator is using a "fair" rule does not directly observe the decision rule used. One standard used is called an outcome test, i.e. comparing, ex-post, the classification assigned by the adjudicator to observed outcomes. For instance, in a criminal justice setting, one may compare the judge's decision to the (somehow obtained) actual innocence or guilt of the defendants, or in a lending setting, compare the lender's decision on whom to extend loans to with the actual repayment outcomes of loan applicants etc.

In this context, a given prescription on what constitutes a "fair" or non-discriminatory rule maps into a corresponding outcome test. In particular, a popular test corresponds to the common-posterior-threshold rule described above which is not the best test in our model. When used, this test attempts to evaluate whether the adjudicator is using a common posterior threshold across groups by evaluating whether the marginal agents across groups have similar probabilities of different outcomes. Implementing this test is difficult: identifying (and being sure that one has correctly identified) the marginal agent in each group is hard. For instance, there may be information observed by the decision maker but not by the regulator. By contrast, if our maintained assumptions are valid, then an adjudicator wishing to minimize crime should use a rule that equalizes false positive and false negative rates across demographic groups. This is easy to estimate and test: there is no need to identify a marginal agent.

ACKNOWLEDGMENTS

Jung, Kannan, Lee, Roth, and Vohra gratefully acknowledge financial support from NSF grant CCF-1763307. Pai gratefully acknowledges financial support from NSF grant CCF-1763349. Extended Abstract. The full paper is available at https://arxiv.org/abs/2002.07147.