Algorithmic Bias: A Counterfactual Perspective*

Bo Cowgill
Columbia University

Catherine Tucker
Massachusetts Institute of Technology

ABSTRACT

We discuss an alternative approach to measuring bias and fairness in machine learning: Counterfactual evaluation. In many practical settings, the alternative to a biased algorithm is not an unbiased one, but another decision method such as another algorithm or human discretion. We discuss statistical techniques necessary for counterfactual comparisons, which enable researchers to quantify relative biases without access to the underlying algorithm or its training data. We close by discussing the usefulness of transparency and interpretability within the counterfactual orientation.

1 INTRODUCTION

The question of algorithmic bias is inherently causal. If a company uses a new algorithm, will the change *cause* outcomes to be more biased or unbiased? If we use one algorithm rather than the status quo, will this *cause* fewer women and minorities to be approved for loans?

The question of algorithmic bias is also inherently *marginal* and *counterfactual*. In many settings, the use of a new algorithm will leave lots of decisions unchanged. A new algorithm for approving loans may select and reject many of the same candidates as a human process. For these *non-marginal* candidates, the practical effect of a new algorithm on bias is zero. To measure an algorithm's impact on bias, researchers must isolate marginal cases where choices would counterfactually change with the choice of selection criteria.

This white-paper discusses quantitative empirical methods to assess the causal impact of new algorithms on fairness outcomes. The goal of total fairness and unbiasedness may be an impossibly high computational and statistical hurdle for any algorithm. Practitioners often need to deploy new algorithms iteratively and measure their incremental effects on fairness and bias, even if the underlying method is an uninterpretable black box. Researchers should focus on interventions that measurably decrease bias and increase fairness incrementally, even if they do not take bias and unfairness to zero. This requires a distinct quantitative toolkit.

Our approach does not require access to the underlying algorithm or its training data, and do not require transparency or interpretability in the usual sense. To the contrary, the methods can be used to study black-box machine learning algorithms and other black-box tools for selection (such as expert judgement or human discretion). We evaluate the merits of interpretability and transparency (as traditionally defined) in our counterfactual orientation.

2 STATISTICS OF COUNTERFACTUAL EVALUATION

Modern causal inference methods were pioneered by Rubin [12], and are popular for measuring the effects of interventions in a variety of fields (medicine, economics, political science, epidemiology and others). In this paper, the introduction of an algorithm is an "intervention" or a "treatment" – akin to a government changing policies or a patient taking a pill – whose causal effects can be measured. A full discussion of these methods are beyond the scope of this two-page whitepaper, but we give a brief overview below.

Suppose we have a choice variable $X \in (0, 1)$. As an example, X can represent whether or not to extend a loan to an applicant. The bank may have two methods deciding how to set X which we can call A and B. Suppose that A is the status quo, and a policymaker must evaluate adopting B. Either A or B could be machine learning algorithms, or both or neither could be. Applicants are indexed by i. X_{Ai} represents whether method A would grant a loan to applicant i, X_{Bi} represents the choice method B would take.

For many practical settings, researchers may find it useful to know what percentage of decisions would change given A vs B, and what covariates were correlated with agreement and disagreement. For example: Do A and B mostly agree on male applicants, but disagree on females? Do they agree on white rejections, but disagree about white acceptances? Measuring the quantity and location of these disagreements will offer early clues about how much a new algorithm will affect racial biases (compared to the status quo).

Even these simple comparisons are missing from most of the current literature on algorithmic bias. For example: In 2016, ProPublica produced an influential analysis [1] of the COMPAS recidivism guidance tool, alleging the COMPAS algorithm was biased against black defendants. Data from this example have been extensively studied in the subsequent academic literature [2, 3, 8].

Even if COMPAS were racially biased, it may not have affected defendants' outcomes. If judges were already predisposed to sentence in a COMPAS-like way – e.g., they agreed independently with COMPAS – the tool would have no effect. This seems likely, given that COMPAS was trained on data that judges could view independently.[1] As we discuss later in this paper, it's possible that a biased algorithm is an improvement upon a status quo counterfactual with greater bias. To our knowledge, only one paper [5] has attempted to measure how often judges (A) and COMPAS (B) would agree or disagree if each made independent evaluations.

Beyond measuring disagreement, causal inference methods offer guidance on establishing which method is right. Suppose we have a "payoff" variable Y. For this exposition, suppose that $Y \in (0, 1)$ representing if the loan is paid back.[2] The choice of Y may reflect inherently subjective policy priorities. For example: If the bank

*This white-paper is contains ideas from a longer article by the authors titled *Algorithmic Bias: Economics and Machine Discrimination* [6]. Bo Cowgill (bo.cowgill@columbia.edu) is an Assistant Professor at the Columbia Business School in New York, NY. Catherine Tucker (cetucker@mit.edu) is the Sloan Distinguished Professor of Management Science at MIT Sloan School of Management, Cambridge, MA, and Research Associate at the NBER.

Working Paper: NSF Trustworthy Algorithms, December 2017, Arlington, VA

[1] In addition, the COMPAS training data incorporated historical judicial behavior as part of the modeled phenomena.

[2] If payment sizes vary, Y could be extended beyond zero and one to give greater payoff when a larger loan is repaid.

valued a diverse loan portfolio, then Y could be coded to give larger payoffs to the bank's utility if minorities receive a loan.

Note that payoffs Y depend on whether the loan is extended (X). We can thus extend the notation to Y_{Ai} and Y_{Bi}, representing the different payoffs depending on which selection algorithm is used.

The problem of evaluating two algorithms is that for many settings, Y_{Bi} is unknown. Historical observational data would contain outcomes only for A (and for where A and B agree). The critical data – what would happen if B overrode A in cases of disagreement – is missing data.

To overcome this problem, the causal inference literature has required researchers to collect new data by finding a random sample of Y_{Bi} outcomes. In our example, we could implement a field experiment overriding A with B randomly, and observing Y_{Bi}. Unlike experiments in other fields, algorithmic evaluation experiments can be relatively easy and safe for the subjects. Researchers generally know what happens if a candidate is not interviewed – the payoff for that candidate to the employer is zero. Thus no experimentation is necessary for rejections.

Instead, a employer simply needs to extend additional interviews to a random set of applicants who would be approved in regime B (but not A) and score these interviews' outcomes. By comparison with many experiments, this is incredibly easy and safe for the subjects and researchers.[3] In addition, researchers studying bias do not need to access the algorithm, its functional form, input variables, numerical weights or training data.

3 RELATED EMPIRICAL LITERATURE

Although causal inference methods are widely used in other disciplines (including most experimental sciences), few empirical computer science papers have used these methods to examine algorithmic bias and fairness. Those that have find *positive effects*, even when the algorithms have been trained on historical data potentially containing bias. [4] studies a field experiment in the use of machine learning for hiring, and finds positive effects on underrepresented groups – including groups underrepresented in the training data.

[9] develops an algorithm for predicting criminal recidivism, and constructs simulated counterfactual outcomes by exploiting the random assignment of judges to cases. The authors' findings "suggest potentially large welfare gains: a policy simulation shows crime can be reduced by up to 24.8% with no change in jailing rates, or jail populations can be reduced by 42.0% with no increase in crime rates." They also show that "the algorithm is a force for racial equity," even though it was trained on historical criminal data.

It may sound counterintuitive that an algorithm may be less biased than the underlying training data. [4] proposes a theoretical model to explain the mechanism for for this outcome. The model shows that even if training data is biased, supervised machine learning can produce *less biased* judgements (if not fully unbiased) than the training data – if the training data exhibits sufficient noise and inconsistency in addition to bias. The noise plays a positive role in decreasing bias by providing positive training examples for underrepresented groups. In this sense, noisy training data is a useful input to machine learning algorithms attempting to decrease bias.

4 TRANSPARENCY AND INTERPRETABILITY

Transparency and interpretability in machine learning has many definitions [7]. To some observers, sharing the data described above would provide helpful transparency about what an algorithm does, and who is affected (compared to the status quo). It would also provide interpretable forecasts of how the new algorithm would *cause* changes in outcomes if it replaced a status quo method.

However, many requests for "transparency and interpretability" ask developers to publish an algorithm's functional form, input variables and numeric weights. However, simply examining code to evaluate and potentially exclude sensitive variables as inputs to algorithms does not guarantee fair, unbiased scoring. Other variables correlated with these sensitive variables – particularly in combination with each other – can still produce a biased evaluation. For example, [11] show examples of algorithms attempting to infer ethnic affinity based on affection for certain cultural products. However, these algorithms ultimately conflate income with ethnic affinity, because income is also correlated with tastes for these products. Similarly, [10] show the tendency of commercial ad-serving algorithms to show STEM job ads to women is distorted by competition from other advertisers bidding for female eyeballs. In both these cases, the source of the distortion is unclear from examining the algorithm's code.

For counterfactual evaluation, transparency and interpretability are less directly helpful. The goal of counterfactual evaluation is to measure how outcome would change under different selection regimes. Knowing details about how method scores candidates doesn't provide insights about the difference between two selection regimes. Even if both new and old regimes are algorithms, rather than the more common case of an algorithm replacing human judgement, then "transparency" may not be directly helpful in evaluating disagreements.

For example, suppose we implement an algorithm that evaluates job applicants to replace human evaluators. [4] contains a real-world example of this: The resume screening algorithm in the paper appeared to give negative weight to candidates from non-elite schools. However, the candidates benefitting from the algorithm included disproportionately non-elite graduates. Human evaluators assessed these credentials even more negatively. These *marginal non-elite graduates* – candidates from non-elite schools who were favored by the algorithm but rejected by humans – performed extremely well in subsequent job-related performance evaluations (better than the average candidate selected by the humans).

The same effect works in the opposite direction. An algorithm that appears to *help* a certain group (based on its numeric weights) might actually have a negative effect on that group. For example: Suppose an algorithm predicts a loan applicant's probability of repaying a loan successfully, and places a strong weight (directly or indirectly) on being African American. If loan officers (or the status quo regime) place higher weight on this attribute, then the

[3] By comparison, subjects in randomized drug trials have to agree to take potentially fatal experimental drugs. Then, they must comply with treatment regimes that are sometimes burdensome to follow. Then, the drug company must wait months or years for health outcomes to be realized. By comparison, the experiments suggested above are must simpler and shield subjects from most of the risk.

introduction of the algorithm may *reduce* minority lending despite the positive weight.

These examples demonstrate how transparency and interpretability provide misleading intuition about the effects of an algorithm.

5 CONCLUSION

Decision-making algorithms arose partly because of the growing availability of cheap, detailed datasets. These same datasets allows researchers, policymakers and businesses to *measure and quantify bias* in an unprecedented way.

Bias in machine learning applications may be a particularly attractive measurement target. These applications arise naturally in data-rich settings, and are codified. However, status-quo, non-algorithmic mechanisms may be equally or more biased, even if they are more difficult to measure, analyze or codify.

A more careful study of these counterfactuals may suggest that few decision methods are truly free of all bias, particularly if guided by historical examples or data. The incumbent framework for regulating discrimination is mostly based on zero bias. These frameworks may discourage adoption of new technologies that reduce bias.

This paper has introduced empirical methods from causal inference for quantifying changes in bias from a new algorithm versus a counterfactual – even if this counterfactual is not itself algorithmic.

The counterfactual orientation of this paper raises new policy questions. If regulators embrace algorithms that reduce (but not eliminate) bias, they may also want to introduce incentives to speed the reduction of bias. What policies encourage algorithm developers to continue prioritizing reducing bias? Without these, algorithm developers may become complacent with incremental improvements on the status quo, when larger decreases in bias may be possible with effort. □

REFERENCES

[1] Angwin, J., J. Larson, S. Mattu, and L. Kirchner (2016). Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica, May 23.*

[2] Chouldechova, A. and M. G'Sell (2017). Fairer and more accurate, but for whom? *arXiv preprint arXiv:1707.00046.*

[3] Corbett-Davies, S., E. Pierson, A. Feller, S. Goel, and A. Huq (2017). Algorithmic decision making and the cost of fairness. *arXiv preprint arXiv:1701.08230.*

[4] Cowgill, B. (2017a). Automating judgement and decisionmaking: Theory and evidence from résumé screening.

[5] Cowgill, B. (2017b). How algorithms impact judicial decisions.

[6] Cowgill, B. and C. Tucker (2017). Algorithmic bias: Economics and machine discrimination.

[7] Doshi-Velez, F. and B. Kim (2017). Towards a rigorous science of interpretable machine learning.

[8] Flores, A. W., K. Bechtel, and C. T. Lowenkamp (2016). False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *Fed. Probation 80*, 38.

[9] Kleinberg, J., H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan (2017). Human decisions and machine predictions. Technical report, National Bureau of Economic Research.

[10] Lambrecht, A. and C. E. Tucker (2016). Algorithmic bias? an empirical study into apparent gender-based discrimination in the display of stem career ads.

[11] Miller, A. and C. Tucker (2017). Algorithms and historical racial bias. *Mimeo, MIT*.

[12] Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology 66*(5), 688.