Measuring the Biases that Matter

The Ethical and Casual Foundations for Measures of Fairness in Algorithms

Bruce Glymour
Department of Philosophy
Kansas State University
Manhattan, KS, USA
glymour@ksu.edu

Jonathan Herington
Department of Philosophy
Kansas State University
Manhattan, KS, USA
jherington@ksu.edu

ABSTRACT

Measures of algorithmic bias can be roughly classified into four categories, distinguished by the conditional probabilistic dependencies to which they are sensitive. First, measures of "procedural bias" diagnose bias when the score returned by an algorithm is probabilistically dependent on a sensitive class variable (e.g. race or sex). Second, measures of "outcome bias" capture probabilistic dependence between class variables and the outcome for each subject (e.g. parole granted or loan denied). Third, measures of "behavior-relative error bias" capture probabilistic dependence between class variables and the algorithmic score, conditional on target behaviors (e.g. recidivism or loan default). Fourth, measures of "score-relative error bias" capture probabilistic dependence between class variables and behavior, conditional on score. Several recent discussions have demonstrated a tradeoff between these different measures of algorithmic bias, and at least one recent paper has suggested conditions under which tradeoffs may be minimized.

In this paper we use the machinery of causal graphical models to show that, under standard assumptions, the underlying causal relations among variables forces some tradeoffs. We delineate a number of normative considerations that are encoded in different measures of bias, with reference to the philosophical literature on the wrongfulness of disparate treatment and disparate impact. While both kinds of error bias are nominally motivated by concern to avoid disparate impact, we argue that consideration of causal structures shows that these measures are better understood as complicated and unreliable measures of procedural biases (i.e. disparate treatment). Moreover, while procedural bias is indicative of disparate treatment, we show that the measure of procedural bias one ought to adopt is dependent on the account of the wrongfulness of disparate treatment one endorses. Finally, given that neither score-relative nor behavior-relative measures of error bias capture the relevant normative considerations, we suggest that error bias proper is best measured by score-based measures of accuracy, such as the Brier score.

FAT '19,* January 29–31, 2019, Atlanta, GA, USA
 ACM ISBN 978-1-4503-6125-5/19/01…$15.00

CONCEPTS

• **Social and professional topics ~ Computing / technology policy** • Theory of computation ~ Design and analysis of algorithms

KEYWORDS

Algorithmic decision-making, fairness, casual inference, discrimination

ACM Reference format:

Bruce Glymour and Jonathan Herington. 2019. Measuring the Biases that Matter: The Ethical Basis for Measures of Fairness in Algorithms. In *Proceedings of ACM Conference on Fairness, Accountability and Transparency (FAT'19).* Atlanta, Georgia, USA https://doi.org/10.1145/3287560.3287573 N

1 INTRODUCTION

Several recent papers [9,11,15,21,36] have argued that various measures of algorithmic bias impose contrary demands on algorithms employed to predict behavior. Here we show how those results fall naturally out of simple considerations of underlying causal structure. By attending to the causal structure, it is possible to sort measures of bias into four types. Each type is motivated by distinct ethical considerations, and is sensitive to different features of the probability distribution over feature variables, variables recording target behaviors, and 'sensitive' variables recording membership in one or another social class. Attention to the underlying causal structure makes clear which constraints on unbiased distributions can and cannot be jointly satisfied, in general and under various relevant special conditions.

We begin by considering the causal structures that make it possible to predict behavior from covariates, turn to a discussion of measures and their joint satisfiability, and close with some normative considerations.

1.1 Causal Structures

There are a number of discussions of algorithmic bias in terms of causation now in the literature [20,22,23]. Here we use the machinery of causal modeling to explore the underlying structures to which distinct measures of bias are sensitive, classify measures with respect to those sensitivities, and note both limits to and tradeoffs among measures of various kinds. Algorithms predict behavior on the basis of observed values of variables that covary with the behavior to be predicted. That covariation, if not due to accident (i.e. sampling error), arises from some underlying causal structure. The machinery of causal graphical modeling permits representations of the possible causal structures, and allows rigorous inferences from them. Here we employ the modeling conventions elaborated in Spirtes [33] (see also Pearl [27]).

Consider an algorithm that takes as input a vector of model variables, and outputs a score representing the (epistemic) probability that the subject being evaluated will engage in some target behavior. For instance, a credit risk rating algorithm might take as input model variables whose values represent a subject's gender, age, employment history, past loans, rental history, etc., and generate as output a score that predicts the chance that the subject will default on a loan, either quantitatively (e.g. as a probability or risk score) or qualitatively (e.g. by assignment to a 'high' or 'low' risk group). Such prediction is possible only because the model variables are associated with the variable encoding the target behavior, and, if the association is not accidental, that in turn requires that the model variables be causally related to the target behavior. Since model variables take their values before the target behavior occurs, this association requires either that: (i) the model variables cause the behavioral variable, or (ii) they share a common cause with the target behavior. These two possible structures are represented in graphs 1a and1b, respectively, where for ease we employ a single model variable M, a single sensitive class variable C, a score variable S, a variable recording the (non)occurrence of the target behavior, B, and a variable O representing some relevant outcome caused by the decision into which the score enters as a consideration (e.g. whether a loan is offered). Arrows in the graphs represent causal dependencies between variables, directed from the cause to the effect. The variable U1 represents some unmeasured and unknown common cause. By convention, the graphs are understood to be 'causally complete', i.e. there are no missing common causes of represented variables and all direct causal (and definitional) relations between variables are represented by arrows in the model.

Hence, the graphs in Figs. 1a and 1b represent situations in which the class variable neither shares a common cause with, nor causally influences nor is causally influenced by any other variable in the graph. That assumption is not always warranted, but when it holds, the class variable will be statistically associated with the score S and the target behavior B only by accident, i.e. as a result of non-representative sampling. When it fails, several possibilities present themselves.

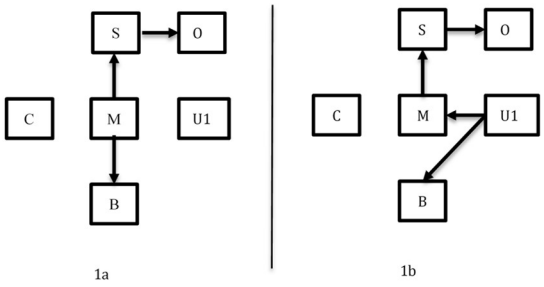

Figure 1: Causal Structures Permitting Prediction

Consider, for the moment, the associations between C and S that are possible when M causes B directly. One causal structure which will generate such an association is that in Fig. 2a. In that graph, C causes S directly, as it will when C is itself among the feature variables employed by the model. Alternatively, C might cause M, as in Fig. 2b. A third possibility is that M causes C, as in Fig. 2c. Finally, M might share a common cause with C, as in Fig. 2d. For some sensitive variables, e.g. those encoding race or sex, it may be that this third possibility can be ignored, on the grounds that the feature in question is necessarily exogenous. But some class variables, e.g. disability status, clearly can be influenced by model variables.

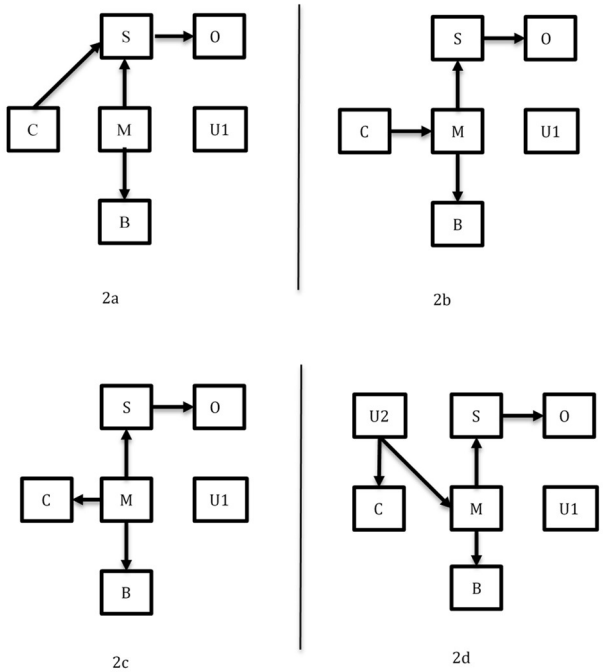

Figure 2: Causal Structures Inducing Associations Between C and S

Turning our attention to an association between C and B, it turns out that when the causal structure is as in Fig. 2a, C will not be associated with B, while when casual structure is as represented in Figs. 2b, 2c or 2d, C will be associated with B. But there are two other ways in which C can come to be associated with B: it may be that C is a direct cause of B, or that they share some common cause (U3), as in Figs. 3a and 3b respectively.

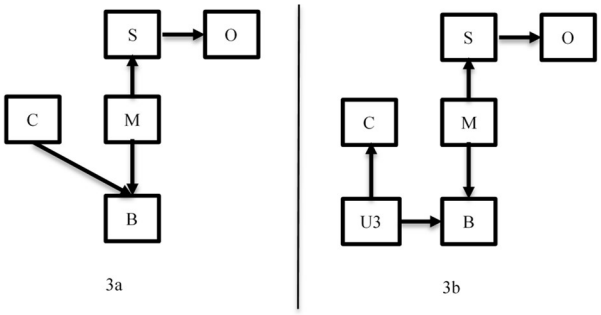

Figure 3: Causal Structures Inducing an Association between Class and Behavior

For the moment, we will assume that the class variable under consideration is exogenous, and hence we will give minimal further consideration to the possibilities represented by Fig. 2c, 2d and 3b.

Things are somewhat different if M is not a direct cause of B, but rather shares a common cause with it. Of special concern is the graph, in Fig. 4a, in which the direct edge between M and B in 2d is replaced by a common cause U1. In that case, C will be associated with S unconditionally, but not with B. We now turn to an explanation of the associations we impute on the basis of the above graphs.

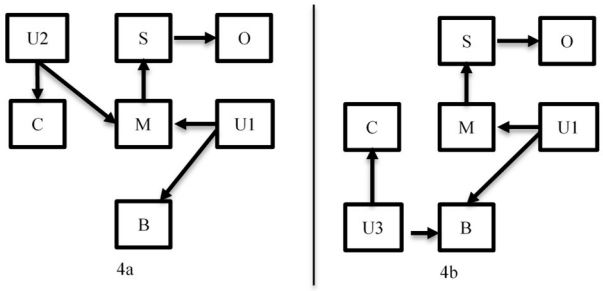

Figure 4: Causal Structures of Special Interest

Associations in observational data are not accidental only if they represent underlying probabilistic dependencies, and those probabilistic dependencies themselves hold in virtue of causal relations among the variables. It is possible to read off from a causal graph which variables are probabilistically dependent and which independent of one another, given a possibly empty set of variables on which one is conditioning. Doing so requires the use of the so-called D-separation Theorem, on the assumption that the Causal Markov and Faithfulness conditions hold (readers are referred to Spirtes [22] for detailed discussion the theorem and the axioms from which it follows). Here we simply recount the rules of thumb by which one may read from a graph facts about which variables will and will not be probabilistically independent from one another [30]. Variables in a path (i.e. a sequence of variables connected by arrows) come in four varieties:

(1) *Terminal* variables, with only one arrow in or out (C and B are terminal variables in the path C→M→B in Fig. 2b);
(2) *Mediators*, which have one arrow in and one arrow out (M is a mediator on the path C→M→B in Fig. 2b);
(3) *Common causes*, which have two arrows out (M is a common cause on the path C←M→B in Fig 2c); and
(4) *Colliders*, which have two arrows in (S is a collider on the path C→S←M→B in Fig 2a).

Two variables connected by a path are associated provided that the path is 'open'. A path is open provided each variable in the path is 'on'. A variable is 'on' in a path, relative to a (possibly empty) conditioning set {V} of variables in the graph, provided it is a terminal variable, a mediator, or a common cause and not in the conditioning set, or is a collider and either is in the conditioning set or has some effect, direct or distal, which is in the conditioning set. If every variable in a path is on, the path is open and the two terminal variables are said to be d-connected on the conditioning set. D-connected variables are probabilistically dependent, conditional on the variables in the conditioning set, and therefore (conditionally) associated in representative data. Variables that are not d-connected relative to a conditioning set are independent of one another, conditional on the variables in the conditioning set.

Using the d-separation theorem, one can see that in Figs. 1a and 1b, C is independent of, and therefore in representative data will be statistically unassociated with S, any outcome O influenced by S, and the target behavior B. There is no path, and thus no open path, between C and S, O or B. Similarly, one can see that in every graph in Fig. 2, there is an open path between C and S: the path is direct, C→S, in Fig 2a; in Fig 2b the path C→M→S includes only terminal variables and a mediator, and thus open when the conditioning set is empty; in Fig. 2c the path C←M→S includes only terminal variables and a common cause and so is also open given an empty conditioning set; and the path C←U→M→S includes U as a common cause and M as a mediator, and so is, unconditionally, open. In Figs. 3a and 3b, there is a direct path C→B is open unconditionally in both graphs; consequently C and B will be associated in representative data given such causal structures.

However, in Fig. 4a, there is only one path between C and B, C←U2→M←U1→B. M is a collider in that path, and so relative to an empty conditioning set, C and B will be unassociated; put formally, C and B are probabilistically independent (C_||_B) in any probability density faithful to Fig. 4a, and in any representative data set they will be statistically unassociated on any measure of association.

2 MEASURES OF BIAS

Measures of algorithmic bias can be classified by the (conditional) dependencies to which they are sensitive. Our critical discussion presupposes that when the dependencies to which a measure is sensitive exist, the measure will report bias.[1] On that assumption, then, one broad set of measures diagnose bias when the score returned by an algorithm is probabilistically dependent on a class variable, like race or sex, either conditionally or unconditionally. In the case that the measure is sensitive to an unconditional dependence between class and score, these measure effectively test for departures from the Independence condition, in the terminology of Barocas and Hardt [3]. Among the various measures which take this form are 'statistical parity' [12] and 'demographic parity' [20]. Other measures in this class test for a conditional dependence between class and score, given the feature variables used by the algorithm (e.g. 'conditional statistical parity' as described by Corbett-Davies et al. [11]). Whether sensitive to conditional or unconditional dependencies between class and score, these measures can be conceptualized as measures of procedural bias, in that they are sensitive to a particular feature of the procedure by which scores are produced. In particular, they detect the direct causal or definitional influence of class (in the conditional case), or some proxy of class (in the unconditional case), on score.

A second group of measures are sensitive to any conditional or unconditional dependence between sensitive class variables and the decisions or outcomes influenced by the algorithm's output, e.g. whether or not a loan or parole is actually granted. Among the various measures which take this form are the '80% rule' [14,26] and the 'Calders-Verwer gap' [26], as well as some definitions of 'statistical parity' which measure outcome disparity rather than score disparity [11]. Other possible measures in this class might test for a conditional dependence between class and outcome, given the score assigned to an individual. Measures in this group can be understood as measures of outcome bias: they detect the causal influence of class on outcome, either through some effect on score (in the unconditional case), or by some other causal process (in the conditional case).

A third group of measures are sensitive to the stability of the relationship between behavior and score when one conditions on class, e.g. measures of 'balance' [21], 'disparate mistreatment' [35], and 'predictive equality'[11]. These measures judge there to be no bias if the probability of receiving a given score, given whether or not one engaged the behavior, is invariant among classes. This is just to require that behavior render class unassociated with score; it is sometimes said of such situations that class is 'screened off' from score by behavior. Measures sensitive to this relationship test for 'Separation' in the terminology of Barocas and Hardt [3]. These measures are often motivated normatively by a concern to avoid scores that are differently reliable for different classes of persons. As such, they might then intuitively be thought of as measures of error bias.

[1] In particular, we assume that the Causal Markov and Faithfulness conditions hold (c.f. Berk [5]).

A fourth group of measures might also be motivated by a concern with differential reliability, and so comprehended in the error bias category, but differ from the third group by a subtle but important formal difference. They also are sensitive to the stability of the relationship between score and behavior conditional on class, but focus on the reverse screening off relation. Such measures – e.g. of calibration [9,21] or predictive parity [9] – judge that no bias occurs when the probability of engaging in the target behavior given one's score is invariant over different classes. If this obtains, class is independent of behavior, conditional on score, i.e. class is screened of from behavior by score. In the terminology of Barocas and Hardt [3], this condition is called 'Sufficiency'.

It is worth emphasizing the important difference between the two classes of measures of error bias. The former class is sensitive to the independence of score and class, conditional on behavior (i.e. whether $S \perp\!\!\!\perp C|B$) while the latter are sensitive to the independence of behavior and class, conditional on score (i.e. whether $B \perp\!\!\!\perp C|S$). We call the first group, including balance, misclassification, predictive equality, and false positive and false negative rates, *behavior-relative* measures of error bias and we call the second group, including calibration and predictive parity, *score-relative* measures of error bias. Intuitively, it might be said that behavior-relative measures are appropriate when one is concerned to ensure that correct and incorrect predictions are similarly distributed in distinct classes, while score-relative measures are appropriate when one is concerned to insure that scores are as informative as possible, no matter what class is being considered. That said, we do not endorse the intuitions or the measures, for reasons given below.

2.1 Causal Structure and Performance

Over the last three years a number of investigators have shown that various measures of bias cannot be driven to zero outside special conditions, and that particular measures of different kinds of bias are jointly incompatible, in that minimizing one requires that one not minimize others [9,11,15,21,36]. And a very few have offered particular measures or methods for constructing measures that are said to avoid such conflict [16,35]. Nearly all of this work proceeds with suppositions that restrict the particular statistical measures of association being employed. Barocas and Hardt [3] is an exception; that discussion provides elegant general proofs that Independence, Separation and Sufficiency are pairwise inconsistent. As it turns out, once one realizes that the various measures of bias are simply different instruments for detecting specific probabilistic dependencies and independencies, these results fall out of simple consideration of the alternative causal structures. Here we report a few such results, some novel and others merely sustaining conclusions reached elsewhere by others on the basis of quantitative rather than structural considerations. As Barocas and Hardt [3] suggest, attending to the underlying causal structure yields insights into what can and cannot be achieved by way of algorithmic fairness, no matter the ingenuity of the design. What is more, careful attention to underlying structure tells us something about when alternative

measures of bias are good or bad measures, depending on the kind of bias with which we are most concerned.

A. Measures, like statistical parity, that are sensitive to a probabilistic relationship between score and class are tests for procedural bias, in its widest sense. Procedural bias can be defined more or less restrictively. On its most restrictive conception, procedural bias occurs only when the class variable is a model variable, directly influencing score (as in Fig. 2a). On slightly less restrictive conceptions, procedural bias also includes cases in which the class variable causes some feature variable included in the model, but which is not itself in the model (as in Fig. 2b). On the most comprehensive conception of procedural bias, all that is required is an association between class as measured by C and score S, and so procedural bias will comprehend cases in which a model variable causes class (Fig. 2c) as well as cases in which class and some model variable share a common cause (Fig. 2d). In each case, there is an open path between C and S, in virtue of which the two are unconditionally associated, and so measures like statistical parity will detect bias. Hence, if one is concerned only with procedural bias in a narrower sense, these are bad measures. But if procedural bias is given its widest interpretation, they are perfectly good measures. Whether a wider or narrow understanding of procedural bias is appropriate as a test of disparate treatment, and so whether extant measures are cogent, depends entirely on the normative considerations at hand. Those are quite likely to vary over contexts, for reasons we briefly report below.

B. Further, procedural bias, on both narrow and wide understandings, is a special case of outcome bias; that is, every case of procedural bias is also a case of outcome bias. This result falls directly out of considerations of causal structure. An unconditional association between class C and score S requires there be an open path between them, and in that path S must be a terminal variable with an edge directed into it. By assumption, score influences outcome. Extending the path between C and S by this edge yields a path between C and O which differs from that between C and S only in that S is now a mediator. Since the path between C and S is unconditionally open, the path between C and O is unconditionally open, and C and O will be associated, i.e. outcome bias will occur in that outcomes will be differentially distributed over classes.

C. Behavior-relative measures of error bias, while normatively motivated by a concern to avoid differential rates of error between classes, are best understood as complex and unreliable measures of procedural bias. Such measures are normatively motivated as follows: score and behavior must be associated, else algorithmic output is a complicated, predictively useless coin flip; but one might desire that this association be invariant among classes, so as to avoid a situation where some groups experience higher rates of being wrongly denied, or wrongly subjected to, the outcome in question. The required invariance holds only when S and C are independent given B. Behavior-relative measures will, therefore, report bias when this conditional independence fails, i.e. when Separation does not hold.

Three problems arise. The first is made explicit by Barocas and Hardt [3] when they show that Independence and Separation are jointly incompatible. Their simple and elegant derivation assumes that C and B are associated unconditionally. For exogenous class variables, that is possible only if class causes behavior. Under that condition, it is easy to see that 'Separation' must fail: conditioning on B will open the path C→B←M→S, inducing an association between C and S, regardless of whether C causes S. 'Independence' may yet be achieved, provided class does not cause any model variable, but 'Separation' is simply not a satisfiable condition. Perforce, Independence cannot be jointly satisfied with either Separation or Sufficiency under these conditions.

Second, conditioning on B will open the paths C←U3→B←M→S in graph 3b and C←U3→B←U1→M→S in graph 4b, inducing an association between C and S when C and B share a common cause. Thus, behavior-relative measures will report bias either when class causes behavior or it shares a common cause with behavior, no matter what other causal relations may or may not exist between class and score (excepting degenerate cases where either score or class is perfectly predicted by behavior).

Third, when C neither causes B nor shares a common cause with it, behavior-relative measures will report bias when, and only when, C causes S, directly or indirectly, or shares a common cause with some model variable, or is an effect of some model variable. But, per point A above, this implies that behavior-relative measures are simply detecting unconditional procedural bias. Moreover, they do so in ways that invite apparent false positive indications of procedural bias. Such false positive indications of procedural bias will arise whenever C causes B but is not causally connected to S by an unconditionally open path (c.f. Chouldechova [9] regarding disparate impact arising from failures of balance). The risk of false positives is eliminated if one simply attends instead to any unconditional association between score and class. Hence, behavior-relative measures are fraught. In effect, they are introduced to detect morally relevant violations of Separation, but actually detect only morally relevant violations of Independence, and do even this unreliably.

D. Score-relative measures of error bias are normatively grounded in a concern to provide individuals with maximally informative predictions that do not take their class membership into account (i.e. the model variables are "sufficient" to accurately predict behavior [3]). Of particular concern is that an algorithm subject to score-relative error bias may be less informative than it could be about one class, while being as informative as it could be about the latter. In that case, it would give some groups less than they deserve, namely optimal prediction given their vector of features. Stipulating the cogency of the motivation, it turns out that these measures too are statistically problematic. To avoid bias as so measured, an algorithm must be constructed so as to insure the independence of B from C, conditional on S (i.e. to ensure Sufficiency). This is not possible if C causes M, or if M causes C, or if C and M share a common cause. As these are exactly the structures that generate

wide sense procedural bias, score-relative measures turn out to be merely complicated measures of disparate treatment, just as with behavior-relative measures. And again, should either class cause the target behavior, or class and behavior share some unmeasured common cause, special concerns arise. When C causes B, C and B will be associated conditional on S, since conditioning on S does not close the path between C and B. That is true even if C is not unconditionally associated with S, i.e. even when disparate treatment is not occurring in any sense. So again, as measures of disparate treatment, score-relative measures of bias are subject to false positive reports that could be avoided by simply attending to the unconditional association between score and class.

E. Error bias of either kind can be avoided only in two ways. First, if C is causally isolated from the variables S, M and B, then C will be unassociated with any of them. In that circumstance, and only in that circumstance, will the frequencies of B and S be non-accidentally invariant over classes. As noted by Chouldechova [9] and Kleinberg [21], it is then possible, in fact, trivial, to ensure joint minimization of score- and behavior-relative measures of bias: C is not d-connected to either S or B on any conditioning set, and so is independent of either variable conditional on the other. Second, if one conditions on a terminal variable in a path, that closes the path. Hence, score-relative measures of bias can be minimized by ensuring that score perfectly predicts class, because in that case to condition on S is in effect to condition on C. And similarly, if it happens that behavior perfectly predicts class or score, behavior-relative measures of bias can be minimized because to condition on behavior is in effect to condition on class (or, respectively, score) (c.f. Kleinberg [21], who reach much the same result quantitatively). Note that perfect prediction of C by S requires the existence of disparate treatment in the wide sense. Thus, score-relative error bias and procedural bias cannot be jointly eliminated, i.e. per Barocas and Hardt, Independence and Sufficiency cannot be jointly satisfied.

F. Sensitive social classes are sensitive in part exactly because they are, or are thought to be, causally implicated in generating behaviors of interest, if only because class conditions society's response to subjects. For example, race is almost certainly a cause not of criminal behavior but of re-arrest, i.e. recidivism-as-measured, if only because police are more likely to arrest African-Americans than members of other racial categories in otherwise similar circumstances. Hence, we should expect that causal connections between class and behavior will be ubiquitous, and error biases of both kinds unavoidable. We suggest this ought to occasion a rethinking of the value of all condition-relative measures of error bias.

3 WHICH BIASES MATTER?

Algorithms serve multiple purposes, and hence are subject to multiple desiderata. One desideratum is that they be as predictively accurate as possible. That requirement is justified broadly by considerations of welfare: the greater the predictive competence of an algorithm, the more efficient the distribution of resources its output informs.

Alongside overall predictive accuracy, we might also care about the fairness of the decisions made by the algorithm. But "fairness" is an essentially contested concept, with a wide range of competing elaborations. In broad terms, we suggest that conceptions of fairness relevant to algorithmic decision-making can be classified in one of three ways, as either: (i) disparate treatment, (ii) disparate impact or (iii) differential distribution of error. In what follows, we discuss the normative considerations that may lead one to care about each form of discrimination, the measures which are apposite to each form of discrimination, and the potential for tradeoffs between such measures.

3.1 Disparate Treatment and Procedure Bias

Paradigmatically, disparate treatment involves an *intention* to disadvantage members of a group by direct sorting according to a class variable: e.g. "whites only" water fountains or 'Irish need not apply' employment ads. Such paradigm cases invite a narrow conception, on which a decision involves disparate treatment if and only if: (i) the decision-maker intends to disadvantage members of a particular class, and (ii) does so by explicitly making his decisions partly on the basis of each subject's membership of that class. But two broader sets of cases also appear to involve disparate treatment.

First, some decisions appear to involve disparate treatment even when they are not intended to disadvantage members of a particular class. For instance, a steel mill that prohibited women from working in the forge, out of a paternalistic concern to benefit the "fairer sex" by protecting them from workplace injury, appears to have engaged in disparate treatment. In such cases, the key ingredient appears to be that decisions are explicitly made on the basis of class membership, regardless of what motivates that usage. Second, some decisions appear to involve disparate treatment even if they do not involve explicit consideration of each subjects' membership of a class. If the foreman at the steel mill instituted a height requirement for employment, because of a desire to exclude as many women as possible from his shop floor, then he appears to have engaged in disparate treatment, regardless of the fact that he does not refer to gender in his employment decisions. And in such a case, the key ingredient appears to be the mere intention to disadvantage women, however it is realized.

Consideration of these non-paradigmatic cases makes clear that disparate treatment involves procedural bias. All of these cases involve an association between class variables and the decision/score. But whether disparate treatment should be associated with a narrow or wide conception of procedural bias requires us to consider why disparate treatment is wrongful.

3.1.1 The Wrongfulness of Disparate Treatment

Judgments about the moral valence of disparate treatment can be grounded in either deontic, motivational or consequentialist considerations. Deontic accounts take unequal treatment to be wrong because it involves treatment based upon morally illegitimate reasons. So, for instance, the judge who denies bail to a black defendant because she is black, even without an intention to disadvantage her, wrongs the defendant because her race is not an appropriate reason for distributing a benefit or a burden.

Precisely why such reasons are illegitimate is contested. On a popular (though probably mistaken [7]) view, such reasons are illegitimate because they involve treatment according to features that are outside of the subject's control [18]. On another view, it is illegitimate because (*ex hypothesi*) race does not cause the defendant's conduct, and thus treats the defendant arbitrarily and without regard to their merit [19]. On a third view, it is illegitimate because the beliefs which control the decision demean people of color: to act on such beliefs is to express them, and to express them is to express the view that people of color are less worthy of just treatment [17,32]. The common refrain amongst these views is that there is something wrong *per se* about class variables directly causing score. Thus, on this account, disparate treatment is identified with the most restrictive definition of procedural bias – i.e. those instances where there is a direct edge between class and score (Fig. 2a). One such measure is "conditional statistical parity" [11], which tests for an association between score and class conditional on the set of "legitimate" model variables. But the utility of this test is limited, since a failure of conditional statistical parity either indicates that (i) there is a direct edge between class and score, or (ii) that there is some unknown model variable that is associated with class. And in the absence of full disclosure of model variables by the developers/users, we cannot infer which of these is the case.

Motivational accounts ground the wrongness of disparate treatment in their connection to an intention to harm or disadvantage the relevant group. By intending to harm the members of a class, the decision-maker reveals themselves to be motivated by animus or ill-will towards that class, in ways which are inconsistent with respecting the personhood of the disadvantaged group [25,29]. And there appear to be multiple ways of realizing such an intention. The history of discrimination shows that the same discriminatory purpose can be achieved by the use of closely correlated proxies for class variables (e.g. zip code, high school graduation rates). Sometimes such proxies will be caused by the sensitive class variable (i.e. if being a woman is a cause of being sexually assaulted, then allowing a history of sexual assault to be a model variable can further intentional discriminatory treatment against women). But proxies are often simply close covariates (i.e. zip code, literacy), chosen without regard to their complex causal relationship to class. In this respect, intentional discrimination can be accomplished by selecting model variables which generate any open path between class and score (i.e. any of Fig 2a-d). Identifying possible cases of intentional disparate treatment thus require taking a maximally expansive definition of procedural bias. If there is any association between class and score, and this association is explained by the decision-makers intent to disadvantage a group, then the decision is an instance of disparate treatment discrimination.

However, in the context of algorithmic decision-making, motivational accounts of disparate treatment face two problems. First, as Binns [6] notes, algorithms do not possess the attitudes of animus, ill-will or intentionality required for direct sorting by class variables to be wrong. While the developers and users of an algorithm can possess those kinds of attitudes, the mere fact that they deploy an algorithm which sorts by class does not establish that they intended that treatment to disadvantage one group. Second, a measure of procedure bias provides only *prima facie* evidence for disparate treatment, one must also provide evidence of the users' motivation. And in so far as algorithms remove much of the decision-making apparatus from the hands of users, they are likely to obscure the kinds of evidence that courts take as establishing intention to discriminate [4].

Consequentialist accounts takes unequal treatment to be wrong because of the downstream effect on the distribution of outcomes [2,24]. In the following section we will explore some reasons one might take an unequal distribution of outcomes across class variables to be of concern. But regardless of the moral grounds, if this is the concern, then class needn't cause score at all. Once again. it is enough that there is simply an association between class and score. Moreover, if we care about disparate treatment because of the resulting outcomes, then disparate treatment ought to be captured by measures of outcome bias, rather than measures of procedural bias.

These foregoing considerations illustrate three important things about our choice of measures of disparate treatment. First, if one is motivated by deontic accounts of the wrongfulness of disparate treatment, then one ought to endorse a very restrictive measure of procedural bias. And current measures of procedural bias, such as "statistical parity" are too broad [11]. Instead, only measures of procedural bias that demonstrate a direct path between C and S, mediated if at all only by model variables should be deployed in investigations of disparate treatment. When C is known to be exogenous, such a path is demonstrated by an association between S and C that is absent when one conditions on the set of all model variables.

Second, if one is attracted to motivational accounts of the wrongfulness of disparate treatment, then no measure of procedural bias is sufficient to identify cases of disparate treatment. While measures of wide-scope procedural bias – such as "statistical parity" – may identify *candidates* for investigation, the crucial ingredient is the intentions and motivations of the decision-maker.

Third, in so far as we care about disparate treatment because of the resulting outcomes, then we ought to care not about procedural bias (i.e. the relation S⫫C), but about outcome bias (i.e. the relation O⫫C).

3.2 Disparate Impact and Outcome Bias

Disparate impact discrimination involves unintentional disadvantaging of members of a group by virtue of a *prima facie* neutral policy or practice. Under U.S. law, it has three elements: (i) a practice causes a disproportionate share of adversity to fall on members of a class, and either (ii) the practice is not necessary to meet the legitimate goals of the decision-maker, or (iii) there is an alternative practice which will result in a less disproportionate distribution of adversity [4]. The first element is measured by unconditional outcome bias – i.e. if there exists any (significant) association between O and C. But the second and third tests

require a more nuanced investigation. In particular, they require us to know something about why disparate impact is unjust.

Judgments about the moral valence of disparate impact can be grounded in either distributional, expressive or desert-based considerations. In the distributional case, we care about disparate impact if and only if it creates or exacerbates unjust patterns in the distribution of goods. So, for instance, a welfare egalitarian might regard a practice which disproportionately harms one group wrongful if it increases overall inequality. For a prioritarian, a practice is wrongful if it makes the least-advantaged group worse off than they would have otherwise been. In the expressive case, we care about disparate impact simply because it results in visible disparities between classes that undermine their equal status as citizens [1,28]. This view appeals to an "anti-caste" principle: that we ought not design social systems such that highly visible distinctions between classes are associated with deprivation [34]. The concern is disparities based upon easily recognizable characteristics (race, gender, age, disability, etc), come to be seen as natural hierarchies. In such circumstances, the social bases for treating one another as political and moral equals is likely to erode.

Importantly, according to distributional and expressive considerations, disparate impact is wrongful when it creates or exacerbates "wide-scope" disadvantages [25]– i.e. if already disadvantaged groups are made worse off by the specific disparity generated by the algorithm. Disparities that are minor, or where the adversity is disproportionately borne by otherwise well-off groups, matter much less (if at all) on these accounts. Moreover, these distributional and expressive goals are threatened by disparate impact regardless of whether the algorithm assists the decision-maker in meeting their legitimate goal. That an algorithm is more predictively accurate, for instance, does not lessen the distributional or expressive impact of the disparities that it generates. The bare disparities in outcome are what matter. In this respect, these considerations may best be captured by an analysis of outcome bias that is sensitive to whether the adversity falls on an already disadvantaged class.

Finally, in the desert-based case, we care about disparate impact because it exacerbates failures to give individuals what they deserve [31]. For instance, where a prima facie neutral practice disproportionately deprives a group who share an unchosen class variable (e.g. sex, race, etc.), it appears that the practice penalizes individuals for the "bad luck" of being born with particular characteristics that are (ex hypothesi) irrelevant to the target behavior. Here, simple consideration of outcome bias is insufficient, and we must instead consider whether the outcome bias is fully explained by reference to the algorithms' predictive accuracy, or whether it is an invidious artifact. Indeed, measures of error bias – such as balance and calibration – are prima facie attractive, precisely because they appear to ask whether similarly risky people are treated equivalently, regardless of class.

3.3 Equal Concern and Error Bias

Condition-relative measures of error bias can, as we noted above, be normatively motivated either by an interest in constraining the difference in error rates between classes, or by appeal to maximizing predictive accuracy without reference to the sensitive attributes. We think on either motivation, such measures turn out to be objectionable.

Behavior-relative measures are initially motivated by the idea that when scores wrongly predict a subject's behavior, generating either false positive predictions that a behavior will occur, or false negative predictions that a behavior will not occur, subjects are wrongly harmed on the basis of their class membership. One explanation of this wrong is that such treatment fails to give the members of the class "equal concern" relative to members of other groups [13:370]. Equal concern, so the thought goes, requires that the same amount of effort is made to accurately classify members of different classes. One interpretation of this is simply that we should minimize predictive error for each class, without reference to the difference in error rates between classes. But then measures of error bias are irrelevant: what matters is overall accuracy. The second interpretation of equal concern is that it requires equality of predictive error across classes. This *is* captured by measures of error bias, but two problems loom.

First, reducing error bias will often require "levelling down" predictive accuracy in one class, in order to equalize with the lower predictive accuracy of the other class. This involves making some individuals in the better predicted class worse off (by failing to predict their behavior as accurately as we could have), in order to satisfy an abstract principle of equality between classes. While it is possible to endorse such a tradeoff, it is notable that no individual of a maligned class benefits *per se* from equality of predictive error (unlike equality of *outcome*, which has distributional or representational benefits).

Second, it is simply impossible to avoid bias, so understood, when class causes behavior. And as the evidence surrounding racialized policing and workplace harassment suggest, we have good reason to suppose that class often causes behavior (e.g. re-arrest, poor performance) precisely because of racial and gender-based injustice. While constructing a *society* where class no longer influences behavior is a worthwhile goal, this does not imply that we should judge algorithmic decision-making as if that goal has already been met. In particular, given that it will be generally impossible to secure homogeneous distributions of error, aiming to avoid these forms of bias may involve reducing the predictive accuracy of algorithms in ways that sabotage efforts to compensate for existing injustices [10]. We do not see why one should prefer algorithms that generate scores less well predicted by class to algorithms that generates scores better predicted by class, (except in the sense that such algorithms exhibit procedural bias, at least in its wide sense).

Score-relative measures are in much the same boat. Grant that an algorithm generating ill-calibrated scores gives persons less than they deserve, in that the algorithm predicts for some less well than it might have done. In so far as we suppose that sensitive attributes are (or should be) irrelevant to our behavior, we might therefore assume that algorithms should be maximally predictively accurate without referring to an individual's class (i.e. the model should be 'sufficient' for prediction). Thus, it is at

least initially plausible that we should aim to construct algorithms which render behavior and class independent conditional on score. But again, that aim can only be notional, because in the case that C causes B (i.e. for almost all cases of interest) the aim is unachievable. Worse, in that case, better predictions of B require that S track C, else C's influence will appear as noise, i.e. error, with respect to the algorithm's predictions. But if S is to track C, it must do so by way of an association induced by an unconditionally open path between C and S, which path will not be blocked by conditioning on behavior. That is, given that sensitive attributes are often (unjustly) relevant to behavior, maximizing predictive accuracy for each class will often require algorithms to include, not exclude, class as a model variable.

Both behavior-relative and score-relative measures can, at least prima facie, be motivated as measures of procedural bias, since measures of both kinds will report the presence of bias when C and S are causally connected directly, indirectly, or by way of a common cause. But so understood, measures of both kinds are seriously unreliable, because they will also report bias when class causes behavior, whether or not class and score are otherwise causally connected. Hence, we think condition-relative measures of bias ought be dispensed with, tout court.

This is not to say that minimizing error is itself not a legitimate aim. It is surely a wrong to mis-predict the behavior of subjects when that behavior could have been predicted by the use of a more accurate algorithm. And equally, less accurate predictions impose a cost in efficiency, thus decreasing overall social utility. But, we suggest, absent special considerations, such error is best assessed by measures which are not condition-relative. Given the inadequacies of condition-relative measures of error bias, we suggest that proper scoring rules used to judge predictive accuracy, e.g. the Brier score [8], might provide better alternative measures of error bias (c.f. Lipton [26] among other discussions employing accuracy based measures of error bias).

Though minimizing error is an important desideratum, it is not, or at least not always, of overriding moral importance, but rather only one consideration among several relevant to a moral assessment of any decision procedure. It is therefore worth noting, with some emphasis, that when class causes behavior, it is simply not possible to avoid procedural bias while minimizing error, and indeed, the most efficient method for minimizing error - i.e. where class is a model variable directly influencing score - involves paradigmatic disparate treatment. Procedural bias and error bias cannot, therefore, be jointly minimized when class causes behavior, and that unfortunate fact cannot be avoided by any statistical measure of bias, no matter how cleverly devised.

4. CONCLUSION

We have examined a number of the causal structures in which class, model, behavior, score and target behaviors may be embedded. Those structures imply a set of conditional and unconditional associations which have a number of implications. In particular, procedural bias, e.g. disparate treatment, turns out to be a special case of outcome bias. Further, when class variables cause target behaviors (and so the behavior is differentially distributed across classes), procedural and condition-relative error bias cannot be jointly avoided. Indeed, under those circumstances, behavior-relative and score-relative bias cannot be jointly avoided. Moreover, condition-relative measures of error bias cannot be given a clear normative motivation that is distinctly different from that warranting attention to procedural and outcome bias (e.g. disparate treatment and disparate impact), but are unreliable indicators of such bias. We suggest therefore that alternative, non-condition-relative, measures of error bias should be considered in their place, recognizing that minimizing error bias and minimizing procedural and outcome bias will necessarily be competing desiderata when class causes behavior.

ACKNOWLEDGMENTS

The authors would like to acknowledge Bill Hsu, Graham Leach-Krouse, Rosa Terlazzo and the seminar audience at the K-State Laboratory for Knowledge Discovery in Databases, for helpful comments and suggestions on early work. This work was supported by a College of Arts and Sciences' Faculty Star Award from the Kansas State University Foundation.

REFERENCES

[1] Elizabeth S. Anderson. 1999. What Is the Point of Equality? *Ethics* 109, 2 (January 1999), 287–337. DOI:https://doi.org/10.1086/et.1999.109.issue-2

[2] Richard Arneson. 2013. Discrimination, Disparate Impact, and Theories of Justice. In *Philosophical Foundations of Discrimination Law*. Oxford University Press, Oxford, 87–113. Retrieved August 22, 2018 from http://dx.doi.org/10.1093/acprof:oso/9780199664313.003.0006

[3] Solon Barocas and Moritz Hardt. 2017. Fairness in Machine Learning. In *Conference on Neural Information Processing Systems, 2017*.

[4] Solon Barocas and Andrew D. Selbst. 2016. Big data's disparate impact. *Calif. L. Rev.* 104, (2016), 671–732.

[5] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2017. Fairness in Criminal Justice Risk Assessments: The State of the Art. *arXiv:1703.09207 [stat]* (March 2017). Retrieved November 2, 2017 from http://arxiv.org/abs/1703.09207

[6] Reuben Binns. 2017. Fairness in Machine Learning: Lessons from Political Philosophy. *arXiv:1712.03586 [cs]* (December 2017). Retrieved July 12, 2018 from http://arxiv.org/abs/1712.03586

[7] Bernard R. Boxill. 1992. *Blacks and Social Justice* (2nd ed.). Rowman & Littlefield Publishers, Lanham, Md.

[8] Glenn W. Brier. 1950. Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.* 78, 1 (January 1950), 1–3. DOI:https://doi.org/10.1175/1520-0493(1950)078

[9] Alexandra Chouldechova. 2016. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. In *arXiv:1610.07524 [cs, stat]*. Retrieved November 7, 2017 from http://arxiv.org/abs/1610.07524

[10] Sam Corbett-Davies and Sharad Goel. 2018. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. *arXiv:1808.00023 [cs]* (July 2018). Retrieved August 22, 2018 from http://arxiv.org/abs/1808.00023

[11] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. *arXiv:1701.08230 [cs, stat]* (January 2017). DOI:https://doi.org/10.1145/3097983.309809

[12] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Rich Zemel. 2011. Fairness Through Awareness. *arXiv:1104.3913 [cs]* (April 2011). Retrieved November 17, 2018 from http://arxiv.org/abs/1104.3913

[13] Ronald Dworkin. 1978. *Taking Rights Seriously: With a New Appendix, a Response to Critics*. Harvard University Press, Cambridge, Mass.

[14] Michael Feldman, Sorelle Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2014. Certifying and removing disparate impact. *arXiv:1412.3756 [cs, stat]* (December 2014). Retrieved November 18, 2018 from http://arxiv.org/abs/1412.3756

[15] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2016. On the (im)possibility of fairness. *arXiv:1609.07236 [cs, stat]*

(September 2016). Retrieved August 22, 2018 from http://arxiv.org/abs/1609.07236

[16] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. *arXiv:1610.02413 [cs]* (October 2016). Retrieved December 4, 2017 from http://arxiv.org/abs/1610.02413

[17] Deborah Hellman. 2008. *When Is Discrimination Wrong?* Harvard University Press, Cambridge, MA.

[18] Richard D. Kahlenberg. 1997. *The Remedy: Class, Race, And Affirmative Action*. Basic Books.

[19] John Kekes. 1993. The Injustice of Strong Affirmative Action. In *Affirmative Action and the University*. Temple University Press, 144–156. Retrieved from http://www.jstor.org/stable/j.ctt14bs9hb.10

[20] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding Discrimination through Causal Reasoning. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett (eds.). Curran Associates, Inc., 656–666. Retrieved November 16, 2018 from http://papers.nips.cc/paper/6668-avoiding-discrimination-through-causal-reasoning.pdf

[21] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *Proceedings of Innovations in Theoretical Computer Science (ITCS)*. Retrieved November 7, 2017 from http://arxiv.org/abs/1609.05807

[22] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual Fairness. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett (eds.). Curran Associates, Inc., 4066–4076. Retrieved November 16, 2018 from http://papers.nips.cc/paper/6995-counterfactual-fairness.pdf

[23] Matt J. Kusner, Chris Russell, Joshua R. Loftus, and Ricardo Silva. 2018. Causal Interventions for Fairness. (June 2018). Retrieved November 16, 2018 from https://arxiv.org/abs/1806.02380

[24] Kasper Lippert-rasmussen. 2006. The badness of discrimination. *Ethic Theory Moral Prac* 9, 2 (April 2006), 167–185. DOI:https://doi.org/10.1007/s10677-006-9014-x

[25] Kasper Lippert-Rasmussen. 2014. Indirect Discrimination is Not Necessarily Unjust. *Journal of Practical Ethics* 2, 2 (2014), 33–57.

[26] Zachary C. Lipton, Alexandra Chouldechova, and Julian McAuley. 2017. Does mitigating ML's impact disparity require treatment disparity? *arXiv:1711.07076 [cs, stat]* (November 2017). Retrieved March 12, 2018 from http://arxiv.org/abs/1711.07076

[27] Judea Pearl. 2009. *Causality: Models, Reasoning and Inference* (2nd ed.). Cambridge University Press, New York.

[28] Philip Pettit. 1999. *Republicanism: a theory of freedom and government*. Oxford University Press, Oxford.

[29] T. M. Scanlon. 2010. *Moral Dimensions: Permissibility, Meaning, Blame* (Reprint edition ed.). Belknap Press, Cambridge, Mass.

[30] Richard Scheines. 1997. An Introduction to Causal Inference. In *Causality in Crisis? Statistical Methods and the Search for Causal Knowledge in the Social Sciences*. University of Notre Dame Press, South Bend, IN. Retrieved from https://www.cmu.edu/dietrich/philosophy/docs/spirtes/notredame.ps

[31] Shlomi Segall. 2012. What's so Bad about Discrimination? *Utilitas* 24, 1 (March 2012), 82–100. DOI:https://doi.org/10.1017/S0953820811000379

[32] Patrick S. Shin. 2009. The Substantive Principle of Equal Treatment. *Legal Theory* 15, 2 (June 2009), 149–172. DOI:https://doi.org/10.1017/S1352325209090090

[33] Peter Spirtes, Clark Glymour, and Richard Scheines. 2001. *Causation, Prediction, and Search* (2nd ed.). MIT Press, Cambridge, Mass.

[34] Cass R. Sunstein. 1994. The Anticaste Principle. *Michigan Law Review* 92, 8 (1994), 2410–2455. DOI:https://doi.org/10.2307/1289999

[35] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. *Proceedings of the 26th International Conference on World Wide Web - WWW '17* (2017), 1171–1180. DOI:https://doi.org/10.1145/3038912.3052660

[36] Indre Zliobaite. 2015. On the relation between accuracy and fairness in binary classification. *arXiv:1505.05723 [cs]* (May 2015). Retrieved August 23, 2018 from http://arxiv.org/abs/1505.05723