# Accuracy and Fairness for Juvenile Justice Risks Assessments

Richard Berk
Department of Criminology
Department of Statistics
University of Pennsylvania

December 29, 2017

**Abstract**

Risk assessment algorithms used in criminal justice settings are often said to introduce "bias". But such charges can conflate an algorithm's performance with bias in the data used to train the algorithm and with bias in the actions undertaken with an algorithm's output. In this paper, algorithms themselves are the focus. Tradeoffs between different kinds of fairness and between fairness and accuracy are illustrated using an algorithmic application to juvenile justice data. Given potential bias in training data, can risk assessment algorithms improve fairness, and if so, with what consequences for accuracy? Although statisticians and computer scientists can documents the tradeoffs, they cannot provide technical solutions that satisfy all fairness and accuracy objectives. In the end, it falls to stakeholders to do the required balancing using legal and legislative procedures, just as it always has.

## 1 Introduction

The recent introduction of "Big Data" and machine learning into the operations of criminal justice institutions has gotten a mixed reception. For some, the promise of decisions both smarter and more fair has led to qualified support (Ridgeway, 2013a,b; Brennan and Oliver, 2013; Doleac and Stevenson, 2016; Ferguson, 2017). "Smart policing" is one instance. For others, the risks of inaccuracy and racial bias dominate any likely benefits (Harcourt, 2007; Starr, 2014; O'Neil, 2016; Angwin et al., 2016). Racial bias inherent in the data used by criminal justice agencies is carried along and magnified by machine learning algorithms; bias in, bias out.

Computer scientists and statisticians have responded with efforts to make algorithmic output more accurate and more fair, despite the prospect of flawed data. Better technology is the answer. But is it? Perhaps even the best algorithms will be overmatched. Perhaps better technology can only go so far. In the end, perhaps the most challenging issues will need to be resolved in the political arena. These are the matters addressed in this paper.

Much past work on algorithmic bias make algorithms the fall guy. But critics are correct that training data can really matter. Moreover, common applications of criminal justice algorithms provide information to human decision makers. It is important to distinguish between that information, human decisions informed by the information, and subsequent actions taken (Kleinberg et al., 2017, Stevensen, 2017). Concerns about *algorithmic* bias properly are raised only about an algorithm's internal machinery.

There is a formal literature on algorithmic accuracy and fairness that can be directly consulted, and good summaries of that literature exist (Berk et al., 2017). However, the expositions can be quite technical and integrating themes are too often lost in mathematical detail. A different expositional strategy is offered here. The issues will be addressed through empirical examples from a dataset rich in accuracy and fairness challenges: predictions of recidivism for juvenile offenders. The real world setting will make the technical content more grounded and accessible. Credible unifying conclusions can then be more easily drawn.

Four broad points will be made. First, there are many kinds of unfairness so that "bias" can materialize along several dimensions. There can be, for example, inequality of treatment, inequality of opportunity, and inequality of outcome. Second, there will be tradeoffs between different kinds of unfairness with some irreconcilable in most real applications. Third, there will be tradeoffs between accuracy and fairness. If an algorithm is designed to be optimally accurate, anything that introduces additional objectives can lead to reduced accuracy. Finally, it is the job of statisticians and computer scientists to document the various tradeoffs in an accessible manner. But the balancing required to address the tradeoffs is not a technical matter. How the tradeoffs will be made is a matter of competing values that will need to be resolved by legal and political action.

# 2 Background

Literature reviews of juvenile criminal justice risk assessments reveal that risk assessments for juveniles and risk assessments for adults raise most of the same issues. (Pew Center on the States, 2011; Vincent et al., 2012; National Institute of Justice and Office of Juvenile Justice and Delinquency Prevention, 2014; Office of Juvenile Research and Delinquency Prevention, 2015). There are concerns about accuracy (Meyers and Schmidt, 2008; Oliver and Stockdale, 2012) and concerns about fairness (Huizinga et al., 2007; Schwalbe, 2008; Thompson and McGrath, 2012). Differences center on the kinds of predictors used and arguably a greater emphasis on determining needs and treatment modalities for juveniles.[1] The discussion to follow centers on the themes of accuracy and fairness.

There is also a small but growing literature addressing more directly the ethnical and legal issues (Hyatt et al., 2001; Tonry, 2014; Ferguson, 2015; Hamilton, 2016; Barocas and Selbst, 2916; Janssan and Kuk, 2016; Zliobaire and Custers, 2016; Kroll et al., 2017). Many important matters are addressed, but not at the level of detail required by this paper. Algorithmic methods have advanced very rapidly. Material to follow will suggest that legal and ethical thinking has fallen behind. The intent is to make important algorithmic details more apparent so that legal scholars are better able to engage.

## 2.1 The Empirical Setting

The empirical setting is a particular state's department of juvenile services. A juvenile enters the system through a "complaint," which most commonly comes from police after an arrest. But parents, teachers, social workers or any citizen may file a complaint. Once a complaint is received, an intake officer determines whether the complaint should be dismissed, whether the individual should be placed under informal supervision, whether diversion to community-based services is appropriate, or whether a petition for court action should be filed. This decision is shaped by procedural requirements

[1] The recent development for adults of "generation four" risk assessments may right the balance (Desmarais and Singh, 2013).

and administrative guidelines filtered through the experiences of each intake officer. Many complaints are resolved without court action.

For any court action, the intake officer recommends whether detention is necessary prior to adjudication. At that point, a juvenile may be placed in community detention, which can include electronic monitoring, day and evening reporting, or private alternative programs. The subsequent adjudicatory hearing determines whether a juvenile is delinquent or in need of supervision. If so, there can be a "sentence" served either at home, under community supervision, in an out-of-home residence, or for those determined to be dangerous to themselves or others, in a secure institution. After release from an institution, there is "aftercare" much like parole for adults.

The decision made by an intake officer can be very challenging and will rest in part on a projection of "future dangerousness." Concerns about violent crimes necessarily are very salient. In principle, the decision could be informed by an actuarial forecast of risk. Accuracy might well be improved over current methods, but there would also be legitimate concerns about fairness. Any forecasting tools should address both. It is important to emphasize that at intake, subsequent actions by a juvenile court or the juvenile services agency cannot be known. Consequently, those actions cannot be used by the intake officer for forecasting. How long an individual will be "under supervision" is also unknown at intake and cannot be used to project risk.

## 2.2 Data

Data were assembled with which to forecast at intake a juvenile's new complaint for a *violent* crime. Included were all juveniles referred to the state's department of juvenile services at least once during calendar year 2006. Some individuals had multiple referrals within this time frame. The first complaint during 2006 (including all associated offenses) was considered the "current" complaint. All offenses before the current complaint were treated as "priors." Any violent offense that took place after the current complaint was defined as a "failure." Violent offenses included assault 1st degree, attempted murder, attempted rape, carjacking, manslaughter, manslaughter by automobile, murder 1st degree, murder 2nd degree, rape 1st degree, rape 2nd degree, and robbery with deadly weapon. Operationally, all failures were referrals to the department of juvenile services or an arrest as an adult. The vast majority of referrals were the product of an arrest.

In this paper, the term "failure" and the term "arrest" are effectively the same. Consistent with the strong preferences of the department of juvenile services, the reference category for an arrest for a violent crime was an arrest for another kind of crime or no arrest at all. Also consistent with their strong preferences, arrests that counted could occur while a juvenile was under the department's supervision as long as that supervision was not detention in a secure institution.

The juvenile offense history file was downloaded at the end of December, 2010. Each individual had up to a 5 year follow-up in the juvenile system. In addition, adult data were downloaded in July, 2011 which extended by one year the follow-up period for arrests as an adult.[2]

Predictors available at intake on a regular basis included all of the following.

1. Race
2. Gender
3. The number of prior escapes
4. The number of prior felony complains
5. The number of prior misdemeanor complaints
6. The number of nonviolent prior complaints
7. The number of violent prior complaints
8. The number of sex-crimes prior complaints
9. The number of prior weapons complaints
10. The number of prior drug-related complaints
11. The total number of prior complaints
12. The age at which there was a first arrest
13. The age at release from supervision

[2] The actual length of the follow-up period depended, therefore, on the age at intake. We use that later as a predictor.

14. An escape as the current complaint
15. A drug-related offense as the current complaint
16. A felony is the current complaint
17. A misdemeanor as the current complaint
18. A violent crime as the current complaint
19. A nonviolent crime as the current complaint
20. A sex crime as the current complaint
21. A weapons crime as the current complaint
22. The total number of current charges

In most settings, there are value-based objections to using race as a predictor. However, race is needed to help reveal and adjust for any race-based unfairness in algorithmic results. Race is not used below when an algorithm is applied but is used to understand the fairness of fitted values and subsequently to increase fairness.[3]

## 2.3 Statistical Methods

The algorithmic procedure used was XGBoost (Chen and Guestrin, 2016). XGboost is a form of gradient boosting drawing heavily on work by Jerome Friedman (2001; 2002), but with a number of clever innovations that increase processing speed enormously. There is also a large number of tuning parameters that with sufficient work can improve performance.

A form of gradient boosting was the method chosen because so much of the literature on fairness builds on risk scores, often in the form of the fitted probabilities for a binary outcome. Random forests would have provided similar accuracy, but no defensible risk scores. Support vector machine would have also forecasted well, but can struggle with large datasets and requires

[3] If race is a powerful predictor, one has a simple example of a fairness-accuracy tradeoff.

selecting a predictor kernel that formally cannot handle categorical predictors like gender and race.[4]

The risk scores produced by XGBoost were used to construct conventional confusion tables. With one exception discussed later, a threshold of .50 was imposed on the fitted probabilities. Values greater than .50 led to an assigned outcome class of an arrest for a violent crime, and values equal to or less than .50 led to an assigned outcome class of no such arrest. Confusion tables were constructed as usual by cross-tabulating the actual outcome class against the assigned outcome class. From the tables, measures of accuracy and fairness were calculated.

Boosting algorithms do not formally converge and in principle, one can boost forever. The number of iterations is usually determined by some measure of fitting error such as the residual deviance. Fitting stops when fitting error appears to stop declining.[5] But this determination is not very precise and substantial overfitting, especially of the risk scores, is a genuine threat. In response, we fit the data for the first analysis with training data and report out-of-sample results using a holdout sample as test data.[6]

We found little evidence of overfitting. Therefore, for all subsequent analyses, we pooled the training and test data to increase the sample size. For each analysis, the fitting error used to determine the number of iterations was the 5-fold cross-validation residual deviance.

[4] Although categorical variables can be coded as indicator variables with a numeric 1 and a numeric 0, those values are arbitrary. Formally, any two numeric values would do. But because of the dot products needed to construct the predictor kernel, the arbitrary values chosen for the indicator variables actually matter.

[5] Because some forms of boosting introduce sampling of the training data (i.e., stochastic gradient boosting), the fitted values contain noise which can make it difficult to determine when a measure of fit is no longer decreasing. The same problem arises if a cross-validation measure of fitting performance is used. In response, some data analysts require that the the lack of improvement continue for several consecutive iterations before the iteration is halted.

[6] The sample of 33,847 observations was randomly split into training data of 20,000 observations and test data of 13,847 observations. There seems to be no commonly accepted approach for determining the relative sizes of the training sample and the test sample although a number of recommendations have been made (Faraway, 2014).

# 3 Accuracy and Fairness in the Empirical Results

Table 1 was constructed from the test data, although the training data yielded nearly the same results. A re-arrest for a violent crime is called a "positive," and the absence of a re-arrest for a violent crime is called a "negative." This is consistent with common practice.

Starting at the first row of the left-most column, the base failure rate is relatively low: the probability of a violent crime arrest is .10. Nevertheless, because of concerns about violence, juvenile justice stakeholders imposed a 5 to 1 costs ratio on the forecasting errors: false negatives were to be treated as five times most costly than false positives. In other words, it was thought to be 5 times worse to release a juvenile who later was arrested for a violent crime than to hold a juvenile who could have been released with no risk. The 5 to 1 target cost ratio was built into the results by the way in which the algorithm was tuned.[7] In practice, it is extremely difficult to arrive empirically at the exact target cost ratio, and in this case, the empirical cost ratio of 5.45 to 1 for false negative to false positives is shown in the second row and first column of Table 1. Cost ratio differences as large as plus or minus 1.0 in this case made no material difference in the results.

| Performance Measure | Full Sample | White Subset | Black Subset |
|---|---|---|---|
| 1. Arrest Base Rate | 0.10 | 0.04 | 0.14 |
| 2. Cost Ratio | 5.45 | 7.07 to 1 | 4.88 to 1 |
| 3. Forecast Arrest | 0.26 | 0.17 | 0.33 |
| 4. False Positive Rate | 0.22 | 0.16 | 0.28 |
| 5. False Negative Rate | 0.40 | 0.17 | 0.36 |
| 6. Arrest Forecasting Error | 0.78 | 0.89 | 0.75 |
| 7. No Arrest Forecasting Error | 0.05 | 0.03 | 0.07 |

Table 1: Test Data Performance Measures For Gradient Boosting Results For A Violent Crime Arrest (Training data N = 20,000, Test data N = 13,487, White N = 6,208, Black N = 7,639)

Calibration is commonly cited as a highly desirable property of risk forecasts (Kleinberg et al., 2016; Chouldechova, 2016). A risk instrument is

[7] For XGBoost, probably the most convenient way to tune for a target cost ratio is by weighting.

calibrated if for a binary outcome the forecasted probability of failure (or success) is the same as the actual probability of failure (or success). In this application, the actual and forecasted probability of a re-arrest for a violent crime should be the same.

However, calibration can be extremely difficult to achieve in practice. One major obstacle is the absence of important predictors in the data. Another major obstacle is measurement error in the predictors or the outcome to be forecasted. In addition, it is effectively impossible to achieve calibration if the cost ratio of false negatives to false positives is not 1.0. Indeed, calibration only makes formal sense when the costs for both kinds of errors are the same.[8]

For these data, Table 1 shows in the third row that the forecasted arrest probability of .26 is much larger than the actual base rate probability of .10. The larger forecasted probability is to be expected because the preferred cost ratio of 5 to 1 makes false positives relatively cheap. In response, the algorithm will favor the less costly false positives over the more costly false negatives. Consequently, there is a larger number of individuals incorrectly forecasted to be re-arrested then if the cost ratio were, say, 1 to 1. This is not an algorithmic error. The gradient boosting procedure properly is responding to expressed stakeholders' policy preferences that trade a relatively large number of false positives to find the true positives.

A false positive rate (i.e., the probability of a false positive) is the probability that when the truth is a negative, the algorithm classifies the case as a positive. Here, that it the probability that an individual is incorrectly classified as a bad risk. From row 4 in Table 1, the false positive rate is .22. Nearly 80% of the time, the algorithm correctly classifies a true negative as such.

A false negative rate (i.e., the probability of a false negative) is the probability that when the truth is a positive, and the algorithm incorrectly classifies the case as a negative. Here, that is the probability the algorithm incorrectly classifies an individual as a good risk. From row 5 in Table 1, the false negative rate is .40. About 60% of the time, the algorithm correctly classifies a true positive as such.

Neither the false positive rate nor the false negative rate speak directly to *forecasting* accuracy. When calculating the false positive and false negative rate, the true outcome in the training data is given, and one learns the

[8] If the cost ratio is not 1.0, the parallel to calibration is that the forecasted costs of classification error is the same as the actual costs of classification error.

chances that an algorithm will find it. False positive and false negative rates primarily are used as metrics to evaluate algorithmic performance within training data. They can provide measures of how well an algorithm performs, which is important to computer scientists and statisticians.

More relevant for policy purposes are forecasting errors. Given that a forecast has been actually made, what are the chances that it is correct? The last two entries in the first column of Table 1 provide that information. When an arrest for a violent crime is forecasted, it has a probability of .78 of being wrong. Although the high probability is disappointing, it results substantially from the 5 to 1 cost ratio discussed earlier. Because on policy grounds stakeholders determined that false positives were far less costly then false negatives, a large number of false positives are folded into forecasts of arrests. Empirically, there are about 4 false positives for every true positive. A smaller cost ratio would have reduced the number of false positives, but a smaller cost ratio would have been inconsistent with stakeholder preferences.

When no arrest for a violent crime is forecasted, the forecast is incorrect with a probably of .05. This is the other effect of the 5 to 1 cost ratio. Because false negatives are so costly compared to false positives, the algorithm works hard to avoid them. The relatively small number of false negatives leads to very high forecasting accuracy when an arrest for a violent crime is not forecasted.

Arrests for violent crimes are rare and difficult to predict. The results from the first column of Table 1 are roughly consistent with past machine learning efforts to forecast crimes that are very troublesome but relatively uncommon (Berk, 2012; Berk and Sorenson, 2016). In this application, considerable success forecasting the *absence* of an arrest for a violent crime is to be expected because the base rate for violent crime is low (i.e., .10). Ignoring all of the predictors, one can be correct about 90% of the time by always forecasting no arrest. It is difficult to do better. But if the algorithmic results are used to forecast the absence of an arrest for violent crime, that forecast would be correct about 95% of the time (row 7). Forecasting error is cut in half. And given the large number of juveniles for whom such forecasts would be made, several hundred young people in a given year would be correctly treated as low risk who otherwise would be incorrectly treated as high risk. The price paid is the relatively large number of violent crime false positives.

But what about fairness? Consider the two columns in Table 1 to the right of the column just discussed. Output from the fitting algorithm was subsetted for White juveniles and Black juveniles, and the same performance measures computed for each.

It is immediately apparent that the base rates are quite different. The probability of an arrest for a violent crime is about .04 for Whites and about .14 for Blacks. It is well known that except in highly stylized illustrations, when base rates differ between groups, various kinds of inequality can cascade through many performance measures, and affect calibration as well (Kleinberg et al., 2016; Chouldechova, 2016; Berk et al., 2017). To take a simple example, an accurate forecasting procedure *should* predict that Black juveniles are more likely to be arrested for a violent crime because that is the truth in the training data. Some would criticize the forecasts as lacking "statistical parity" or lacking "demographic parity." But trying to fix a lack of statistical parity can have undesirable side effects (Dwork et al., 2012). For example, one could fix the problem by detaining a sufficiently large, random sample of lower risk, White juveniles. Clearly, this would be a policy non-starter.

In Table 1, there are other fairness concerns as well.

1. The cost ratios differ. For Whites the cost ratio is a little over 7 to 1. For Blacks the cost ratio is a little under 5 to 1. For Whites, it is more important than for Blacks to avoid assigning a low risk label to juveniles who are actually high risk. Black juveniles may be getting something like a beneficial algorithmic thumb on the scale. In the last two rows, one can see that this can lead to greater forecasting accuracy for *Whites* when a prediction of no violent crime arrest is made. (i.e., The *number* of false negatives is fewer.) At the same time, this can lead to greater forecasting accuracy for *Blacks* when a prediction of a violent crime arrest is made. (i.e., The *number* of false positives is fewer.) How does one reconcile these different forecasting outcomes with existing conceptions of fairness? One suggests algorithmic bias against Blacks and one suggests algorithmic bias against Whites.

2. From row 3, one can see that calibration is achieved for neither group. As noted earlier, this disparity results from Blacks having a larger base rate to begin with. Should the Black-White disparity be seen as unfair nevertheless, one solution might be to require stronger statistical evidence for Blacks than whites to forecast an arrest for a violent crime. One might try to justify this strategy by arguing that the higher base rate for Blacks is itself a symptom of unfairness and needs to be dis-

counted. However, one would be introducing *differential treatment* for Blacks and Whites, which some might claim is unfair.

3. From rows 4 and 5 in Table 1, one can see that both false positive rates and false negative rates are higher for Blacks.[9] By this metric, the gradient boosting algorithm performs better for Whites, which some would identify as a lack of equal opportunity (Hardt et al., 2016). However, from a policy perspective, this is an an intermediate concern. No decisions are made from the false positive or false negative rate because in practice when they are computed, the outcome necessarily is known. If the outcome is known, there is no need to forecast it. What matters more is algorithmic output that directly affects decisions at intake when the outcome is not known.

In summary, many fairness and accuracy concerns are raised by the results reported in Table 1. It should be apparent that there are several kinds of unfairness in tension with one another. Accuracy could be better as well. What might be done?

# 4 Proposed Solutions

There are several kinds of unfairness and an absence of calibration that are apparent from Table 1. To date, none of the proposed remedies provide an across-the-board solution (Pedreschi et al., 2008; Kamiran and Calders, 2009; Kamishima et al., 2011; 2012; Hajianm and Domingo-Ferrer, 2015; Feldman et al., 2015; Chouldechova, 2016; Joseph et al., 2016; Hardt et al., 2016; Johnson et al., 2016; Berk et al., 2017; Calmon et al., 2017, Corbette-Davies, 2017; Johndrow and Lum, 2017; Ridgeway and Berk, 2017). Rather, researchers concentrate on one or two kinds of unfairness sometimes coupled with calibration. Then, there are proofs demonstrating that except for some highly unrealistic scenarios, one cannot simultaneously have calibration and equal predictive accuracy. (Kleinberg et al., 2016; Chouldechova, 2016).

One can organize the technical literature using conceptual distinctions from Romei and Ruggieri (2014). Some proposals favor "pre-processing" methods in which adjustments are made to the data before a forecasting

[9] The number of false positives and false negatives must be distinguished from the rate of false positives and negatives.

algorithm is trained. The intent is to remove sources of unfairness from data before the data are used. Other proposals offer different kinds "in-processing" in which the algorithm itself is altered to increase fairness in outputs. And some favor "post-processing" in which forecasting results are manipulated so that greater fairness is produced. But all, at least implicitly, trade one kind of fairness for others in ways that can undermine forecasting accuracy. For example, efforts to produce equality of outcome (e.g., the probability of arrest is the same for black and whites) typically introduce inequality of treatment (e.g., prior arrests for one of the protected classes are given less weight) and dilute predictor information. One can also easily arrive at adjustments that make everyone equally worse off. For example, forecasting accuracy is made the same across all classes, but less accurate on the average for everyone.

It is impractical in this paper to demonstrate how even a small subset of the proposed solutions are supposed work. Far too much exposition would be required. But it is relatively easy to show empirically what the proposed solutions hope to accomplish, key levers they will pull, and the limits of what is likely to be possible.

## 4.1 Equalizing Cost Ratios

Suppose for our data, equal cost ratios are imposed for Black juveniles and White juveniles. One might start by altering the cost ratios by protected class because cost ratios can be determined by stakeholders before a forecasting algorithm is applied. Stakeholders have the option, if they wish, of imposing identical cost ratios in service of greater fairness. Such cost ratios can be important for algorithmic performance because they affect the relative representation of false negatives and false positives in the results that, in turn, affect equality of treatment and equality of output.

Using a form of "in-processing" to this end, separate forecasting exercises were undertaken for White Juveniles and Black Juveniles tuned to arrive at the same target cost ratio of 5 to 1. That is, the algorithm was applied to the White juveniles by themselves and to the Black juveniles by themselves. The full dataset was used because there was no evidence earlier of overfitting and more precise results would be produced. Table 2 shows the results.

One might argue with considerable justification that by applying the gradient boosting algorithm separately to White and Black juveniles, one has implemented in-processing introducing treatment inequality not present before. And in fact, using measures of variable importance, there were some

| Performance Measure | White Subset | Black Subset |
|---|---|---|
| 1.Arrest Base Rate | 0.04 | 0.14 |
| 2. Cost Ratio | 4.83 to 1 | 4.67 to 1 |
| 3. Forecast An Arrest | 0.10 | 0.29 |
| 4. False Positive Rate | 0.08 | 0.22 |
| 5. False Negative Rate | 0.38 | 0.30 |
| 6. Arrest Forecasting Error | 0.75 | 0.66 |
| 7. No Arrest Forecasting Error | 0.02 | 0.06 |

Table 2: Performance Measures From Gradient Boosting Results For A Violent Crime Arrest Separately for White and Black Juveniles With Equivalent Cost Ratios By Race (White N = 15,804, Black N = 18,763)

differences by race in which predictors were driving the forecasts.[10] For example, the four predictors in order making the greatest contribution to the fitted values for Whites juveniles were (1) the number of prior misdemeanors, (2) the age of an offender's earliest arrest, (3) the offender's age when released, and (4) the number of weapons priors. For Black juveniles, the four predictors in order were (1) the number of weapons priors, (2) the number of priors for drug offenses, (3) gender, and (4) the age when released. If the predictors that matter most differ for Whites and Blacks, by this criterion the algorithm is treating White juveniles and Black juveniles differently.

Table 2 shows the new results in detail. The relative weight of false negatives to false positives is now nearly the same for Whites and Blacks: 4.83 versus 4.67 respectively. One important kind of unfairness has been effectively eliminated, but by and large, the earlier kinds of inequality remain. For example, there is still a large gap by race in the forecasted probability of an arrest for a violent crime.

There are apparently no formal methods that favor any version of this cost-ratio approach, perhaps because differences in cost ratios have not been salient fairness concerns. Cost ratios can matter a great deal to stakeholders because, as noted earlier, different kinds of classification errors have different real-world consequences unrelated to fairness, at least as usually defined.

[10] Predictor importance is measured by the average over iterations of contribution to the fit. Details can be found in Hastie et al., 2009: Section 15.3.2.

## 4.2 Equalizing Base Rates

Different base rates for different protected classes have long been understood as important sources of algorithmic unfairness. It follows that a potential pre-processing approach would alter the violent arrest base rates. One can do this by weighting the data. Because the base rate for Black juveniles is about 3 times the base rate for White juveniles, arrests for Black juveniles were discounted by a factor of 3 when algorithm was applied. If one assumes that White juveniles are the “privileged” class, the Black base rate can be weighted to closely approximate the White base rate. Blacks are given the same re-arrest base rate as Whites. Table 3 shows some real improvements in fairness.[11]

| Performance Measure | White Subset | Black Subset |
|---|---|---|
| 1. Arrest Base Rate | 0.04 | 0.04 |
| 2. Cost Ratio | 1 to 30.6 | 1 to 8.1 |
| 3. Forecast An Arrest | 0.002 | 0.007 |
| 4. False Positive Rate | 0.001 | 0.005 |
| 5. False Negative Rate | 0.98 | 0.92 |
| 6. Arrest Forecasting Error | 0.59 | 0.59 |
| 7. No Arrest Forecasting Error | 0.04 | 0.04 |

Table 3: Training Data Performance Measures For Gradient Boosting Results For A Violent Crime Arrest With Weighting (White N = 15,804, Black N = 18,763)

The forecasted probability of an arrest for a violent crime, the false negative rate, the false positive rate and both kinds of forecasting errors are approximately the same for Black juveniles and and White juveniles. As intended, the algorithmic output for Black juveniles now looks a lot like the output for White juveniles. Even for algorithm skeptics, this might be a very satisfactory outcome.[12]

[11] Implementing weighting can be tricky because of the way an algorithm is designed. For example, weighting may be implemented at the fitting stage, but not at the stage when predictions from test data are undertaken. It can be a good strategy to alter the base rates in the training data and test data before applying the algorithm rather than counting on the algorithm getting it right. The procedure is then a mix of pre-processing and in-processing.

[12] The numbers are rounded. With more decimal places, small differences appear.

But there is strong evidence of inequality of treatment. The cost ratio for Black juveniles is quite different from the cost ratio for White juveniles (1 to 8.1 versus 1 to 30.6 respectively). Relative to false positives, false negatives are less costly for Blacks. This is one consequence of discounting by a factor of 3 arrests of Black juveniles because it is the same is increasing by a factor of 3 the relative number of black juveniles who are not arrested. With many more arrest-free Black juveniles, there will be, other things equal, a larger number of false positives.

In short, inequality of treatment is introduced so that one can better approximate equality of outcome. Whether this is "fair" overall will depend in part of why Black juveniles have a substantially higher base rate to begin with. If the higher base rates are substantially and demonstrably caused by bias in the criminal justice system, one might argue that compensating for that bias is appropriate. By itself, however, this ignores the impact that violent crime can have on victims. One might have to show that a higher base rate for violent crimes committed by Black juveniles is primarily an artifact of bias and that the recorded arrests, by and large, do not correspondent to actual victims of violence. Were the arrests for violent crimes typically linked to real crime victims, a far more complicated set of tradeoffs must be considered, perhaps especially because the most likely victims would also be Black. Such concerns are very unusual in the formal work on algorithmic methods. The algorithm proposed by Corbette-Davies and his colleagues (2017) is one exception.

Nevertheless, trying to correct for the difference in base rates can in many situations be well worth trying, and there are several defensible ways one can proceed. The weighting approach is easy to apply almost regardless of the algorithm, but lacks much formal structure. There are more elegant approaches. For example, Kamiran and Calders (2009) gradually change the base rates for two protected classes while not changing the overall base rate over both groups. The average base rate over groups, not the base rate of the more "privileged" group, becomes the common base rate. Their algorithm proceeds by changing the fewest possible actual class labels from success to failure or from failure to success to arrive at equal base rates. This can be a more satisfying way to alter the base rate, but it must confront many of the same limitations. One has introduced treatment inequality. The group with the lower base rate will tend to have its members "downgraded" to failures. The group with the higher base rate will tend to have its members "upgraded" to successes.

One does not have to be limited to making the base rate race-neutral. One can try to extract racial content from all predictors by capitalizing on the usual covariance adjustments in regression as a form of pre-processing (Berk, 2008). Each legitimate predictor is regressed in turn on a problematic variable such as race. From each regression, the residuals are computed as usual. Any linear dependence between each predictor and race is removed because the residuals are by construction uncorrelated with the race variable. One can then proceed using the residuals (but not race) as predictors in the fitting algorithm. Unfortunately, associations with race can remain because in the residualizing process, interaction effects between the legitimate predictors and race are not taken into account. Those include not just the usual two variable interactions, but higher order interactions as well. The solution would be to include all interaction effects, some of which would be of very high order, in the residualizing regressions, but that would increase the number of predictors enormously and introduce disastrous multicollinearity. Far more sophisticated versions of the residualizing strategy have been proposed (Johnson et al., 2016; Johndrow and Lum, 2017), but both still depend on getting the residualizing model right and even then, only some kinds of unfairness are addressed.

## 4.3 Altering the Probability Thresholds

Post-processing also has potential. As noted earlier, many algorithmic methods output a risk score, often interpreted as a probability. Standard practice imposes a threshold of .50 on those probabilities. Individuals with risk scores greater than .50 are assigned to the high risk class. Individuals with risk scores equal to or less than .50 are assigned to the low risk class. The value of .50 is chosen in part because it represents the value for which each outcome is equally likely. But other thresholds can be used that can differ by protected class.[13]

Using the full dataset, gradient boosting was employed as before. The target cost ratio was still 5 to 1, and the empirical cost ratio was 5.6 to 1. The only material change was that a different threshold strategy was imposed for Black juveniles and White juveniles. Table 4 shows the algorithmic output for the Black and White subsets of cases.

[13] Risk scores do not have to be probabilities or even bounded at 0.0 and 1.0. But if forecasted classes are desirable, a threshold of some kind is typically imposed.

The White threshold was maintained at the conventional value of .50. Those with fitted probabilities greater than .50 were projected to fail through an arrest for a violent crime. Those with fitted probabilities equal to or less than .50 were projected to not fail through an arrest for a violent crime. For Blacks, the threshold value was set at .73 so that nearly the same fraction of White juveniles and Black juveniles were forecasted to fail. Outcome equality was achieved using unequal treatment. For both Black juveniles and White juveniles, the fitted probability of an arrest for a violent crime is .11. But as Table 4 shows, there is more to the story.

| Performance Measure | White Subset | Black Subset |
|---|---|---|
| 1. Arrest Base Rate | 0.04 | 0.14 |
| 2. Cost Ratio | 5.25 to 1 | 1 to 1.85 |
| 3. Forecast An Arrest | 0.11 | 0.11 |
| 4. False Positive Rate | 0.09 | 0.04 |
| 5. False Negative Rate | 0.40 | 0.50 |
| 6. Arrest Forecasting Error | 0.78 | 0.34 |
| 7. No Arrest Forecasting Error | 0.02 | 0.08 |

Table 4: Training Data Performance Measures For Gradient Boosting Results For A Violent Crime Arrest With Different Thresholds for White and Black Juveniles (White N = 15,804, Black N = 18,763)

By altering the classification threshold, a dramatic change in the cost ratio was introduced. For Whites, the false negatives are 5.25 times more costly than false positives. For Blacks, false positives are 1.85 time more costly than false negatives. This is another example of treatment inequality. However, there is improved racial equality for the false positive and false negative *rates* and for forecasting error associated with an absence of an arrest for a violent crime. The forecasting error rate for a violent crime arrest still shows a substantial disparity.

There are many approaches to post-processing that adopt a similar strategy, but in a far more elegant manner. For example, Corbett-Davies and his colleagues (2017) build a constrained optimization approach that uses different race-specific risk thresholds to trade public safety against different kinds of fairness. The result is important insights about the potential magnitude of different in-processing tradeoffs, but as the authors note, everything depends on how well their utility maximization model captures actual risk in real settings. In addition, their effort is largely about making the accuracy-fairness tradeoffs more apparent, not figuring out ways to be accurate and fair at the same time. Alternatively, rather than simply shifting the classification threshold, Hardt and his colleagues (2016) choose random cases that have their assigned class reassigned. This approach is placed in an optimization framework so that not just those cases near the threshold of .50 have their forecasted outcome altered while the false positive rates for the different protected classes are made the same (i.e., an equal opportunity constraint).

## 4.4 Regularization Methods

Fairness regularization is a method in which an algorithm's code is hand-tailored to adjust for far more subtle unfairness concerns. More is involved than minor modifications to existing software. The basic idea is to introduce unfairness costs into the fitting process that the algorithm then tries to avoid. Drawing heavily on earlier work (Kamashima et al., 2011; Berk et al., 2017c), Ridgeway and Berk (2017) derive an in-process regularized form of gradient boosting with two regularization terms. One is the usual function that penalizes undesirable complexity in the fitted values. There are mathematical incentives preventing the algorithm from capitalizing on unimportant patterns in the data. The other is a function that penalizes a particular kind of unfairness. There are mathematical incentives for results to be more fair.

Suppose as before that there are two protected classes, say, Whites and Blacks. Each Black individual is compared one by one to all White individuals. When the *actual* outcome for the Black case and the White case is the same (e.g., both were arrested) but the *fitted* scores used for classification are not (e.g., the predicted probability of an arrest differs), there is an instance of unfairness. The greater the disparity in the fitted scores, the greater the unfairness. The algorithm does not care about the direction of the unfairness. Any unfairness may favor either the Black or White individual – the algorithm is formally race-neutral.

The same steps are then undertaken in which each White individual is compared one by one to all Black individuals. Again, the algorithm is formally race-neutral. The two sources of unfairness are then combined to get an overall unfairness measure. With tuning, the impact of overall unfairness can be varied. As one increases the algorithmic incentives for fairness, accuracy declines so that one precisely can document the fairness-accuracy tradeoffs.[14]

The fairness regularizer can be seen as a form of weighting. With each iteration, cases subject to greater amounts of unfairness receive greater emphasis by the algorithm. The algorithm works harder to fit the cases with more unfairness, which can reduce the unfairness. There is not one set of weights but many because the weights are revised with each iteration of the algorithm. The fairness weighting implemented earlier is done only once.

# 5 Conclusions and Recommendations

Summarizing the many details, there are five conclusions. First, it can be useful to think of the different kinds of unfairness as special cases of outcome inequality or treatment inequality. These correspond to conventional categories in jurisprudence. But it remains to be seen whether the two kinds of inequality are sufficiently refined to be useful in an algorithmic context. For example, for different protected groups, one can in service of fair outcomes introduce different weights, different thresholds, different cost ratios, or entirely different algorithmic results. These are all forms of treatment inequality. Are there no important legal distinctions between them? Moreover, there needs to be greater clarity about what an outcome is. There is the information produced by an algorithm, there are decisions that might be made with that information, and there are actions taken as a consequence of those decisions. An algorithm can only be held accountable for the information it provides. The responsibilities for what is done with that information lie elsewhere. In what sense are algorithms ever responsible for unfair outcomes? If there are to be liabilities, who has them?

Second, recall that calibration requires that the predicted probability for a given outcome is the same as the actual probability of that outcome. Although calibration is highly desirable in principle, it is unlikely to be obtained in practice, except in extremely unusual or stylized settings. One has to question, therefore, whether calibration is worth legal analysis. But in the absence of calibration, one might still consider the legal relevance of statistical parity – whether the projected probability of failure (or success) is the same for all protected classes. Perhaps more important is whether the algorithmic fore-

[14] The authors recognize that there are many defensible ways to define a fairness regularizer and that the algorithm can be (and perhaps should be) rewritten so that other manifestations of bias can be addressed.

casts are equally accurate across protected classes. In short, concerns about algorithmic accuracy can come in at least three different forms.

Third, past efforts to increase fairness typically ignore a critical issue: what exactly is the target for equality? This concern was raised earlier when adjustments were made for protected class base rates. What base rate should be the target? The White base rate? The Black base rate? Something between? The same issues arose when different failure thresholds were chosen for different protected classes. Recall that the goal was to have equal predicted probabilities of an arrest for a violent crime. But what predicted probability should be the target? In both cases, the target was determined by the more privileged class (i.e., Whites). There is no mathematical or statistical justification for this choice. The choice was implicitly justified by a view that Whites are privileged, that such privilege is undesirable, and it is the job of the algorithm to fix it by proceeding as if Blacks were equally privileged. One could certainly challenge these views in many ways. The larger point is that considerations of fairness must provide detail on what the fairness target should be and why.

Fourth, discussions of fairness in criminal justice settings typically ignore fairness for victims. This seems myopic in part because perpetrators tend to victimize people like themselves. The leading cause of death among young African-American males is homicide. The most likely perpetrators are other young African-American males. When one adjusts algorithms to make them more fair for Black perpetrators, one risks increasing unfairness for Black crime victims. The importance of their victimization can be discounted and even ignored.

Finally, there are complicated tradeoffs between different kinds of fairness and between different kinds of fairness and different kinds of accuracy. You can't have it all. Computer scientists and statisticians will over time provide far greater clarity about these tradeoffs, but they cannot be (and should not be) asked to actually make those tradeoffs. The tradeoffs must be made by stakeholders through legal and political processes. This will be very challenging.

With all of the unresolved issues, there is currently no preferred way to make the tradeoffs. Stakeholders in each setting need to determine how best to proceed, and there should be room for a variety of different arrangements. At the moment, among the greatest obstacles are negotiations that are poorly informed. Computer scientists and statisticians must provide information on the choices available in an accessible form, and stakeholders must be prepared to listen and learn. Another important obstacle is stakeholders who refuse to acknowledge that compromises are required. Sometimes strongly held values do not play well with facts. A substantial stalemate, or even misguided policies, may be necessary prerequisites for meaningful compromise.

# References

Angwin, J, Larson, J., Mattu, S., and Kirchner, L. (2016) “Machine Bias” https://www.propublica.org/article/machine-bias-risk-assessments -in-criminal-sentencing

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

Berk, R.A. (2008) “The Role of Race in Forecasts of Violent Crime.” *Race and Social Problems* 1: 231 – 242.

Berk, R.A. (2012) *Criminal Justice Forecasts of Risk: A Machine Learning Approach*. New York: Springer.

Berk, R.A., and Hyatt, J. (2015) “Machine Learning Forecasts of Risk to Inform Sentencing Decisions.” *The Federal Sentencing Reporter* 27(4): 222 – 228.

Berk, R.A., Heidari, H., Jabbari, Kearns, M., Morganstern, J., Neel, S., and Roth, A. (2017b) “A Convex Framework for Fair Regression.” *arXiv” 1706.02409v1 [cs.LG]*

Berk, R.A., Heidari, H., Jabbari, Kearns, M., and A. Roth. (2017c) “Fairness in Criminal Justice Risk Assessments: The State of the Art.” arXiv:1703.09207v2 [stat.ML].

Breiman, L. (2001) “Random Forests.” *Machine Learning* 45: 5–32.

Brennan, T., and Oliver, W.L. (2013) “The Emergence of Machine Learning Techniques in Criminology.” *Criminology and Public Policy* 12(3): 551 – 562.

Calmon, F.P., Wei, D., Ramamurthy, K.N., Varshney, K.R., (2017) “Optimizing Data Pre-Processing for Discrimination Prevention,” arXiv: 1704.03354v1 [stat.ML].

Chen, T., and Guestrin, C. (2016) “XGBoost: A Scaleable Tree Boosting System.” arXiv:submit/1502704 [cs.LG].

Chouldechova, A. (2016) “Fair Prediction With Disparate Impact: A Study of Bias in Recidivism Prediction Instruments.” arXiv:1610.075254v1 [stat.AP]

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Hug, A. (2017) “Algorithmic Decision Making and Cost of Fairness.” *KDD ’17: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

Demuth, S. (2003) “Racial and Ethnic Differences in Pretrial Release Decisions and Outcomes: A Comparison of Hispanic, Black and White Felony Arrestees.” *Criminology* 41: 873 – 908.

Desmarais, S.L., and Singh, J.P. (2013) *Assessing Recidivism Risk: A Review of Validation Studies Conducted in the U.S.* New York: Council of State Governments Justice Center..

Dieterich, W., Mendoza, C., Brennan, T. (2016) “COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity.” Northpoint Inc.

Doleac, J, and Stevenson, M. (2106) “Are Criminal Justice Risk Assessment Scores Racist?” Brookings Institute. https://www.brookings.edu/blog/up-front/2016/08/22/are-criminal-risk-assessment-scores-racist/

Dwork, C., Hardt, Y., Pitassi, T., Reingold, O., and Zemel, R. (2012) “Fairness Through Awareness.” In *Proceedings of the 3rd Innovations of Theoretical Computer Science*: 214 – 226.

Faraway, J.J. (2014) “Does Data Splitting Improve Prediction?” *Statistics and Computing* 26 (1-2): 49 – 60.

Feldman, M., Friedler, S., Moeller, J., Scheidegger, C., and Venkatasubrtamanian, S. (2015) “Certifying and Removing Disparate Impact.” In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 259 – 268.

Friedler, S.A., Scheidegger, C., and Venkatasubramanian, S. (2016) “On The (Im)possibility of Fairness).” axXiv1609.07236v1 [cs.CY].

Friedman, J. (2001) “Greedy Function Approximation: A Gradient Boosting Machine.” *Annals of Statistics* 29(5): 1189 – 1232.

Friedman, J. (2002) “Stochastic Gradient Boosting.” *Computational Statistics & Data Analysis* 38(4): 367– 378.

Ferguson, A.G. (2105) “Big Data and Predictive Reasonable Suspicion.” *University of Pennsylvania Law Review* 163(2): 339 – 410.

Ferguson, A.G. (2017) *The Rise of Big Data Policing: Surveillance, Race, and the Future of Law Enforcement* New York: NewYork University Press.

Hajianm S., and Domingo-Ferrer (2013) “A Methodology for Direct and Indirect Discrimination Prevention in Data Mining. *IEEE transactions on knowledge and data engineering*, 25(7):1445 – 1459.

Harcourt, B.W. (2007) *Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age.* Chicago, University of Chicago Press.

Hardt, M., Price, E., Srebro, N. (2016) “Equality of Opportunity in Supervised Learning.” In D.D. Lee, Sugiyama, U.V. Luxburg, I. Guyon, and R. Garnett (eds.) *Equality of Opportunity in Supervised Learning.* Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain, (pp.3315 – 3323).

Hastie, T., Tibshirani, R., and Friedman, J. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, second edition. New York: Springer.

Hamilton, M. (2016) “Risk-Needs Assessment: Constitutional and Ethical Challenges.” *American Criminal Law Review* 52(2): 231 – 292.

Huizinga, D., Thornberry, D., Knight, K., and Lovegrove, P. (2007) *Disproportionate Minority Contact in the Juvenile Justice System: A Study of Differential Minority Arrest/Referral to Court in Three Cities.* Washington, D.C.: U.S. Department of Justice, Office of Justice Programs, OJJDP.

Hyatt, J.M., Chanenson, L. and Bergstrom, M.H. (2011) “Reform in Motion: The Promise and Profiles of Incorporating Risk Assessments and Cost-Benefit Analysis into Pennsylvania Sentencing.” *Duquesne Law Review* 49(4): 707 – 749.

Janssan, M., and Kuk, G. (2016) “The Challenges and Limits of Big Data Algorithms in Technocratic Governance.” *Government Information quarterly* 33: 371 – 377.

Johndrow, J.E., and Lum, K. (2017) “An Algorithm for Removing Sensitive Information: Application to Race-Independent Recidivism Prediction.” arXIV:1703.049557v1 [Stat.AP].

Johnson, K.D., Foster, D.P. and Stine, R.A. (2016) “Impartial Predictive Modeling: Ensuring Fairness in Arbitrary Models.” arXIV:1606.00528v1 [stat.ME].

Joseph, M., Kearns, M., Morgenstern, J.H., and Roth, A. (2016) In D.D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (eds.) *Fairness in Learning: Classic and Contextual Bandits.*Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain (pp. 325 – 333.

Kamiran, F., and Calders, T. (2009) “Classifying Without Discrimination.” *2009 2nd International Conference on Computer, Control and Communication*, IC4 2009.

Kamiran, F., and Calders, T. (2012) “Data Preprocessing Techniques for Classification Without Discrimination.” *Knowledge Information Systems* 33:1 - 33.

Kamiran, F., Karim, A., and Zhang, X. (2012) “Decision Theory for Discrimination-Aware Classification.” IEEE 12th International Conference on Data Mining.

Kamishima, T., Akaho, S., and Sakuma, J. (2011) “Fairness-aware Learning Through a Regularization Approach.” Proceedings of the 3rd IEEE International Workshop on Privacy Aspects of Data Mining.

Kleinberg, J., Mullainathan, S., and Raghavan, M. (2016) "Inherent Trade-Offs in Fair Determination of Risk Scores." arXiv: 1609.05807v1 [cs.LG].

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig., and Mullainathan S. (2017) "Human Decisions and Machine Predictions." NBER Working paper 23180. National Bureau of Economic Resaerch.

Kroll, J.A., Huey, J., Barocas, S., Felten, E.W., Reidenberg, J.R., Robinson, D.G., and Yu, H. (2017) "Accountable Algorithms." *University of Pennsylvania Law Review*, forthcoming.

Liu, Y.Y., Yang, M., Ramsay, M., Li, X.S., and Cold, J.W. (2011) "A Comparison of Logistic Regression, Classification and Regression Trees, and Neutral Networks Model in Predicting Violent Re-Offending." *Journal of Quantitative Criminology* 27: 547 – 573.

Meyers, J.R., and Schmidt, F. (2008) "Predictive Validity of the Structured Assessment for Violence Risk in Youth With Juvenile Offenders." *Criminal Justice and Behavior* 35(3): 344 – 355.

National Science and Technology Council (2016) "Preparing for the Future of Artificial Intelligence." Executive of the President, National Science and Technology Council, Committee on Technology.

National Institute of Justice and Office of Juvenile Justice and Delinquency Prevention (2014) *Prediction and Risk/Needs Assessment.* Justice Research. Washington, D.C.

Office of Juvenile Research and Delinquency Prevention (2015) "Literature Review: Risk/Needs Assessments for Youth." https://www.ojjdp.gov/mpg/litreviews/RiskandNeeds.pdf

O'Neil, C. (2016) *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* Crown Publishers.

Oliver, M.E., and Stockdale, K.C. (2012) "Short- and Long-Term Prediction of Recidivism Using the Youth Level of Service/Case Management Inventory in a Sample of Serious Young Offenders." *Law and Human Behavior* 36(4): 331 – 44.

Pedreschi, D., Ruggieri, S, and Turini, F. (2008) "Discrimination-Aware Data Mining." KDD2008, August 24 – 27, 2008, Las Vegas, Nevada, USA.

Pew Center of the States, Public Safety Performance Project (2011) "Risk/Needs Assessment 101: Science Reveals New Tools to Manage Offenders." The Pew Center of the States.www.pewcenteronthestates.org/publicsafety.

Ridgeway, G. (2013a) "The Pitfalls of Prediction." *NIJ Journal* 271.

Ridgeway, G. (2013b) "Linking Prediction to Prevention." *Criminology and Public Policy* 12(3) 545 – 550.

Ridgeway, G., and Berk, R. (2017) "Fair Gradient Boosting." Working paper, Department of Criminology, University of Pennsylvania.

Romei, A. and Ruggieri, S. (2014): "A Multidisciplinary Survey On Discrimination Analysis." *Knowledge Engineering Review* 29(5): 582 – 638.

Salman, J., Coz, E,.L., and Johnson, E. (2016) "Florida's Broken Sentencing System. Sarasota Herald Tribune. http://projects.heraldtribune.com/bias/sentencing/

Schwalbe, C.S. (2008) "A Meta-Analysis of Juvenile Risk Assessment Instruments." *Criminal Justice And Behavior* 35(11): 1367 – 1381.

Silver, E., & Chow-Martin, L. (2002) "A Multiple Models Approach to Assessing Recidivism Risk: Implications for Judicial Decision Making." *Criminal Justice and Behavior* 29: 538 – 569.

Starr, S.B. (2014b) "Evidence-Based Sentencing and The Scientific Rationalization of Discrimination." *Stanford Law Review* 66: 803 – 872.

Stevensen, M. (2017) "Assessing Risk Assessments in Action." Working paper, Antonin Scalia School of Law, George Mason University.

Thompson, A.P., and McGrath, A. (2012) "Subgroup Differences and Implications for Contemporary Risk-Need Assessment with Juvenile Offenders." *Law and Human Behavior* 36(4): 345 –55.

Tonry, M. (2014) "Legal and Ethical Issues in The Prediction of Recidivism." *Federal Sentencing Reporter* 26(3): 167 – 176.

Vincent, G.M., Guy, L.S., and Grosso, T. (2012) *Risk Assessment in Juvenile Justice: A Guidebook to implementation.* New York. N.Y. Models for Change.

Zliobaite, I., and Custers, B. (2016) "Using Sensitive Personal Data May Be Necessary for Avoiding Discrimination in Data-Driven Decision Models." *Artificial Intelligence and the Law* 24(2): 183 – 201.