Layers of Bias:

A Unified Approach for Understanding Problems with Risk Assessment

Laurel Eckhouse

University of Denver

Kristian Lum

Human Rights Data Analysis Group

Cynthia Conti-Cook

The Legal Aid Society (New York City)

Julie Ciccolini

The Legal Aid Society (New York City)

Author Note

Laurel Eckhouse, Department of Political Science, University of Denver

Kristian Lum, Lead Statistician, Human Rights Data Analysis Group

Cynthia Conti-Cook and Julie Ciccolini, Staff Attorney and Paralegal, The Legal Aid Society (New York City)

The authors thank the Human Rights Data Analysis Group, the National Science Foundation Graduate Research Fellowship Program, the Horowitz Foundation for Social Policy, and the Travers Department of Political Science at UC Berkeley for support for this research. We also thank Patrick Ball, Josh Norkin, William Isaac, and Christopher Shea for valuable feedback. We also thank Emily Salisbury, Jody Sundt, and Breanna Boppre, as well as two anonymous reviewers.

Correspondence concerning this article should be addressed to Laurel Eckhouse at laurel.eckhouse@du.edu.

Abstract

Scholars in several fields, including quantitative methodologists, legal scholars, and theoretically-oriented criminologists, have launched robust debates about the fairness of quantitative risk assessment. As the Supreme Court considers addressing constitutional questions on the issue, we propose a framework for understanding the relationships among these debates: layers of bias. In the top layer, we identify challenges to fairness within the risk assessment models themselves. We explain types of statistical fairness and the tradeoffs between them. The second layer covers biases embedded in data. Using data from a racially biased criminal justice system can lead to unmeasurable biases in both risk scores and outcome measures. The final layer engages conceptual problems with risk models: is it fair to make criminal justice decisions about individuals based on groups? We show that each layer depends on the layers below it: without assurances about the foundational layers, the fairness of the top layers is irrelevant.

Layers of Bias: A Unified Approach for Understanding Problems with Risk Assessment

When someone is accused of a crime, should they be held in jail or released to await trial? Once someone has been convicted, should they be sentenced to imprisonment, or released on parole or probation? In most states, judges make these decisions based on presentations by counsel for the prosecution and defense, either in writing or orally (or both). Laws, enacted in statutes and passed down through judicial precedent, govern what factors counsel may argue and what judges may lawfully consider. For example, the statute that governs what factors New York judges consider does not include the future dangerousness of the defendant (N.Y. Criminal Procedure Law § 510.30) whereas in other states, it does (Gouldin, 2016). Even in one state, laws governing different decision-making moments in the criminal justice process may be different according to the liberty at stake and evidence available for each different decision. Judges will also be constrained in what factors they may consider based on fundamental constitutional principles of equal protection and due process.

Judges, whether appointed or elected, have generally erred on the side of incarceration (Gouldin, 2016). More recently, however, a growing criminal justice reform movement asks judges to rely less on incarceration (Gouldin, 2016). This leads to difficult questions about which defendants should be detained while awaiting trial, which people convicted of crimes should be imprisoned, and for how long. Historically, judges alone have had authoritative discretion to evaluate the suitability of pretrial release or probation, constrained only by the risk of an appellate court reversing them for abusing their discretion, and by constitutional constraints on fundamental fairness.

These constraints are meant to prevent implicit and explicit bias from driving judges' decisions, but there is plenty of evidence that human decision-making is infected by bias (Banks,

Eberhardt, & Ross, 2006; Eberhardt, Davies, Purdie-Vaughns, & Johnson, 2006; Kang, Bennett, Carbado, & Casey, 2011; Levinson, 2007; Levinson, Cai, & Young, 2010; Richardson & Goff, 2012). For this reason, judges, parole boards, and other criminal justice decision-makers (as well as immigration authorities) are increasingly turning to data-driven models to predict who will reoffend. Vendors promote these models to the public and to the agencies that use them as the answer to human bias, arguing that computers cannot harbor personal animus or individual prejudice against anyone based on race, gender, or any other legally protected characteristic.

Data-driven risk assessment tools have become commonplace in the criminal justice system: from where to deploy police resources to who should be imprisoned or granted probation or parole, computer models are increasingly being used to inform—and sometimes determine—decisions. Early versions of risk assessment instruments have been around since the 1920s, but Virginia was the first state to officially adopt one in 1994. Originally designed to improve on human judgment by offering structured techniques for decision-making, risk assessment tools have grown in complexity (Andrews, Bonta, & Wormith, 2006). Many now use the "risk-needs-responsivity" (RNR) model, though the "Good Lives Model" has been proposed as an alternative strategy (Andrews, Bonta, & Wormith, 2011; Polaschek, 2012). These models endeavor to combine assessments of risk with referrals for treatment and support (Andrews, Bonta, & Wormith, 2011; Polaschek, 2012).

In recent years, the use of risk assessment tools has expanded dramatically. In particular, they have come to play a key role in decisions about sentencing, probation, and pre-trial detention. Over twenty state courts use some form of risk assessment at sentencing and a single tool, the Arnold Foundation's Public Safety Assessment, is in use throughout three states, as well as in more than two dozen local jurisdictions (Schuppe, 2017). Actuarial tools also form the core

empirical basis for the "risk-need-responsivity" model of treatment and rehabilitation (Bonta & Andrews, 2007). Similarly, the NCCD's "Structured Decision-Making" model is used to guide decision-making across several seemingly disjoint areas including child protective services, foster care placement, juvenile justice, adult protection, and welfare-to-work services (Children's Research Center, 2008). As a whole, these models provide a set of rules and instruments for standardizing the allocation of government resources and services across a wide variety of settings (Freitag & Park, 2008).

In this article, we focus on criminal justice decisions mostly in the pre-trial release and sentencing contexts, because we believe the stakes are highest when liberty decisions are involved. However, the same conceptual issues apply to any model or quantitative prediction of risk. A full history of the development of risk assessment tools is beyond the scope of this paper, but the interested reader can refer to Monahan and Skeem (2016) or Andrews, Bonta, and Wormith (2006) for valuable overviews.

Data-driven models offer judges and policymakers the appearance of objectivity. These models generate scores for risk of flight, risk of re-arrest, risk of parole violations, and other public concerns, based on data from other people with similar characteristics to the defendant. With the scores as guidance, judges and policymakers can simply apply the same model to every case, and claim that they have used an objective, neutral mechanism which offers fair treatment.

Researchers across disciplines, though, have serious concerns about the biases encoded in these models. In one of the most publicized examples, ProPublica released a report alleging signs of racial bias with a sentencing model called COMPAS, developed by Northpointe, Inc. (Angwin, Larson, Mattu, & Kirchner, 2016). Northpointe responded by arguing that, using a different criterion, there was no indication of racial bias in the model (Flores, Bechtel, &

Lowenkamp, 2016). This controversy extended an existing debate among computer scientists, statisticians, and social scientists about what constitutes bias and fairness in these models and how to define and evaluate it (Corbett-Davies, Pierson, Feller, Goel, & Huq, 2017; Feldman, Friedler, Moeller, Scheidegger, & Venkatasubramanian, 2015; Kleinberg, Mullainathan, & Raghavan, 2016; Kroll et al., 2016). However, as we argue in this paper, there are both practical and conceptual problems facing even an algorithm that has been declared "fair."

Scholars and decision makers need tools to assess data-driven models of criminal behavior. Faced with a dizzying array of technical critiques—and the reality that most judges and policymakers have little statistical background—we need a unified approach to evaluating algorithmic decision-making.

A Unified Approach

This article proposes a unified approach to assessing algorithmic fairness, conceptualized as three hierarchical layers. We describe the pitfalls and concerns in each layer and suggest policy options to improve researchers' ability to evaluate the fairness of models at each level of analysis. The key problem we engage is how to think about and evaluate data-driven risk assessment. While our results focus on risk assessment in the criminal justice system, the method we propose also applies to the dozens of other problems where data-driven decision-making tools are now in use: credit and lending decisions, hiring, advertising, etc.

We approach this as both researchers and practitioners, treating this article as an act of public scholarship. Two of the authors are practitioners at New York Legal Aid Society. As practitioners, we interact frequently with situations where risk assessment is used to determine our clients' risk of flight or rearrest. Our central concern is whether the data being used to make that assessment is reliable, or instead a result of historical policing disparities in targeted

communities, and whether our clients are being falsely classified as "high" risk, especially on the basis of constitutionally protected classes like race, gender, or ethnicity. The other two authors are scholars with quantitative backgrounds (a statistician and a political scientist), who are concerned about the quality of data and analysis being used to support quantitative claims. This article is intended to be broadly accessible and to provide a unifying framework for evaluating risk assessment instruments.

First, we provide a simple, non-technical explanation of how risk assessment models work. This hypothetical example provides a motivating case from which to unpack the main conceptual problems with using data-driven assessment strategies. Then, we develop the *layers* framework to conceptualize the problems with algorithms as belonging to three layers. Each layer depends on the layers below it: without assurances about the fairness of the foundational layers, the fairness of the top layers is irrelevant.

Understanding Models

Let us focus on a frequent prediction question: will someone who has committed one crime commit another in the future? Emily is a hypothetical defendant pleading guilty to robbery.[1] Our hypothetical Judge Adams has to decide if Emily should be offered probation or sentenced to time in prison. Judge Adams wants to know: if she releases Emily on probation, is Emily likely to be arrested again, presumably for committing another crime? Judge Adams may look to a predictive model to help answer the question.

A predictive model starts with many examples of people—potentially millions—who

[1] Emily's name is drawn from a famous study of racial inequality in hiring (Lavergne & Mullainathan, 2004).

have already been through the criminal justice system. A simple model might divide these millions of people into various categories based on their characteristics. Emily's category might be very specific. For example, Emily might be an employed, college educated woman between 25 and 30 who has been arrested at least once before. Models vary in their complexity, but a wide variety of individual traits, characteristics, and personal beliefs may be included (Oleson, 2011; Starr, 2015).

Once Emily's data points are entered into the model, it finds all the people with data points "like Emily" from the potentially millions of data points it has been given. Looking at all the people "like Emily", the model compares how many people were arrested while on probation and how many were not. The percentage of people "like Emily" who were arrested while on probation becomes the model's prediction for how likely it is that Emily will be arrested on probation. For example, if 60% of the people "like Emily" were arrested while on probation within one year, the model might translate that to the judge as a "risk score" that Emily would be "high risk" for rearrest. If a defense attorney can analyze the data points in the model, he or she might argue that Emily is not similar to the people "like Emily". However, what information the model considers about Emily may be proprietary, so the defense attorney may not be able to defend Emily against the score. Based on the prediction, Judge Adams may well find that Emily is too likely to be rearrested to be given a chance on probation.

Real models are far more complicated and sophisticated than this simple example. Nevertheless, it demonstrates the essential pieces of how computer models work and the assumptions they are based on. Every model includes a set of data and a way to predict outcomes from characteristics.

Understanding Layers of Bias

Emily's story raises three types of questions. First, is the model that assigns her a risk score fair? This has been the central point of contention in most of the coverage of risk assessment models thus far and forms the top layer of fairness considerations (Angwin et al., 2016; Chouldechova, 2016; Corbett-Davies et al., 2017; Dieterich, Mendoza, & Brennan, 2016; Kleinberg et al., 2016; Kroll et al., 2016; deMichele et al., 2018). To decide if the algorithm—the method that translates those data points into Emily's score—is fair, we need a statistical definition of fairness. As we will see in the following section, this is not simple. Indeed, different definitions of fairness are often fundamentally incompatible: resolving bias in one way produces a different type of bias.

Second, are the data used to calculate Emily's score biased in some fundamental way? This is the middle layer: the biases embedded in the data. Biased data is a fundamental problem in criminal justice. Because *both* the data used to calculate the risk score *and* the data used to evaluate its success suffer from the same types of bias, that bias is highly unlikely to be correctable without access to outside data.[2]

Finally, is it fair to use data about *other people* to make decisions about Emily's liberty? All models fundamentally rely on the assumption that we can use the behavior of other people to

[2] Both the top and middle layer raise questions about defendants' access to information used to make sentencing and liberty decisions. Often, both the model and the data it's made from are considered proprietary information or trade secrets (Wexler, 2017). Occasionally the scoring formula is released (for example, the Arnold Foundation publicizes the methods by which its risk scores are calculated) but even then the underlying data and the methods for developing the risk score are typically kept secret from the court and public.

decide whether a particular defendant is too dangerous to release. This is the base layer: should we use data on groups to make decisions about individuals? This is in part a legal question, which the courts have not yet adjudicated.

Each layer depends, conceptually, on the ones below it. If making judgments about individuals based on groups is unfair or illegitimate, the quality of the data and the models do not matter. If the data are biased, an apparently fair model merely reproduces that bias. In order to develop a usable, fair model for criminal justice risk assessments, we need to believe that all three layers are fair.

In this paper, we explain how to assess algorithmic bias in each layer. We describe the specific concerns raised by each layer, evaluate the available tools for measuring and mitigating algorithmic bias, and discuss those concerns in relation to specific examples of quantitative risk assessment tools. We conclude by discussing the implications of the use of algorithms in the criminal justice system. We do not provide a comprehensive treatment of the benefits and costs of such algorithms: rather than offering a judgment on the value of specific algorithms, our goal is to provide a unified framework for evaluating them.

Top Layer: Fair Algorithms

Defining Fairness

Defining fairness is not a straightforward proposition in any domain and statistics is no different. There is a robust body of work trying to define various concepts of fairness in the contexts of statistical analysis and risk assessment in particular. This is what we call the top layer of the debate: does the risk assessment model make fair predictions? There are three basic questions that go to whether groups are treated fairly.

First, does the score generated by a model mean the same thing across different groups? For example, if we take all the people who, based on the model, have a 30% chance of committing another crime in one year, about 30% of them should in fact commit a crime. In a fair test, the value of that prediction should be similar across protected groups. So, if a tenth of black defendants with a "30%" score commit a crime, while half of white defendants with the same score commit a crime, the test does not exhibit predictive fairness. Kleinberg et al. (2016) call this "calibration within groups", while Chouldechova (2017) calls it "test fairness."[3]

Second, do people who go on to *not* commit a crime get similar scores across groups? So, for example, maybe black defendants who *do not* go on to commit more crimes are much more likely to get high-risk scores than white defendants who *do not* go on to commit more crimes. In that case, black defendants who turn out to be "safe" get treated more harshly than white defendants who have the same outcomes. Kleinberg et al. (2016) call this "balance for the negative class", while Chouldechova (2017), using the more usual term, calls it the "false positive rate."

Finally, do people who *do* go on to commit crimes get similar scores across groups? That is, if we look back on defendants of different races (or genders, etc.) who *did* commit a crime, did they receive similar risk scores? If, for example, white defendants who turn out to commit a crime look less risky up front and are therefore less likely to be denied bail, white defendants who *do* ultimately commit a crime might be more likely to make bail than black defendants who likewise ultimately commit a second crime. Kleinberg et al. (2016) call this "balance for the

[3] Risk assessment instruments typically translate the probability of reoffense into an ordinal scale rather than a probability.

positive class", while Chouldechova (2017) again uses the more usual term, calling it the "false negative rate."

Achieving fairness on all three measures is not just practically difficult: it is in fact conceptually impossible if there are differences in the measured rate of reoffending across different groups (Chouldechova, 2007; Kleinberg et al., 2016).[4] This impossibility theorem is crucial to understanding the obstacles to developing a fair model: it is mathematically impossible to develop a model that will be "well-calibrated" in the sense of having equal predictive value across groups, *and* fair in the sense of treating members of groups similarly in retrospect.

The COMPAS debate. A specific example helps to unpack this problem, though predictive models of all types face similar issues.[5] In 2016, ProPublica reporters analyzed a risk assessment instrument known as COMPAS, the Correctional Offender Management Profiling for Alternative Sanctions, designed by the for-profit company Northpointe (Angwin et al., 2016). The COMPAS model uses multiple variables and a proprietary (i.e. secret and unverifiable) algorithm to come up with the probability of rearrest for each defendant. Predictions about black defendants, ProPublica argued, systematically overstated the risk those black defendants posed. In fact, ProPublica found that of those who were not rearrested, 45% of black defendants had been flagged as high risk. By comparison, only 23% of white defendants who were not rearrested

[4] As we'll see in the following section, there can be differences in the measured rate of reoffending even when behavior is identical across groups.

[5] For an analysis of predictive validity in the National Council on Crime and Delinquency's risk assessment tool, used to support its structured decision-making program, see Schwalbe et al. (2006).

were flagged as high risk. This speaks to the second question raised above: are people who do not reoffend treated similarly across groups?

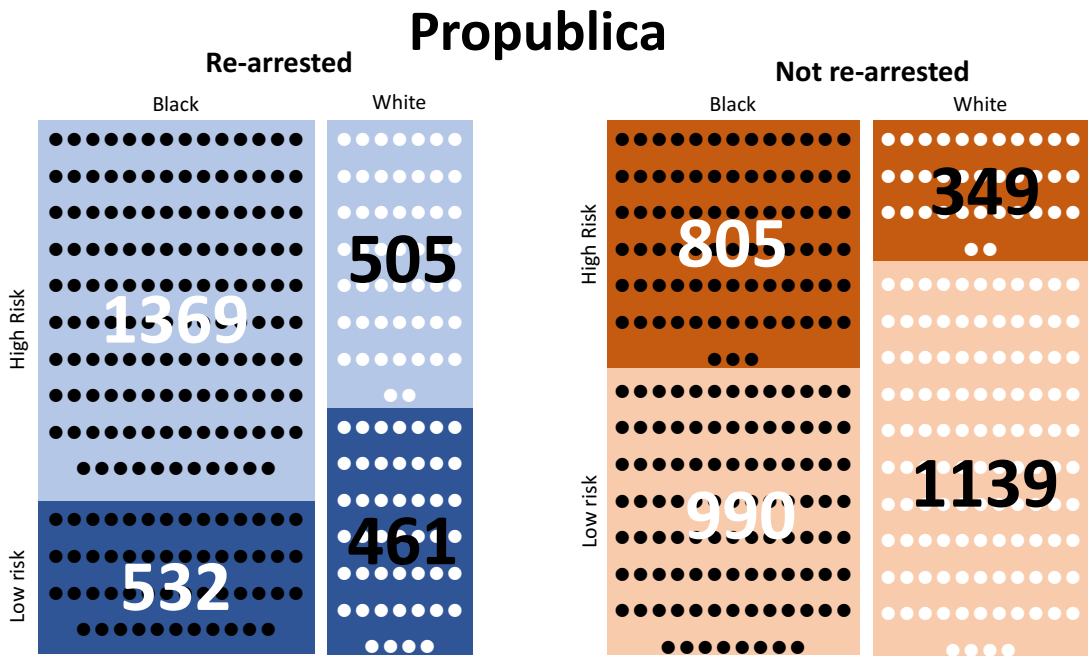

Figure 1. ProPublica's Analysis: categorize defendants by whether they went on to be rearrested. There are clear differences in the "riskiness" categorization for black and white defendants with the same outcomes, as shown by the unequal height of the bars in this graph.

ProPublica came to this conclusion by using the formulation in Figure 1, in which each dot represents 10 people. In the simplest terms, ProPublica worked backward from the outcome. ProPublica first grouped all defendants by the actual outcome, asking whether defendants identified as "high risk" were actually rearrested after being released on probation. They then compared the proportion of "high risk" defendants who were not rearrested.

By analyzing the accuracy of the predictions by race, ProPublica concluded that COMPAS was nearly twice as likely to inaccurately predict that a black defendant was at high

risk for rearrest as a white defendant. The visualization here shows that of those who were rearrested, white defendants were more likely to have been classified as low risk than black defendants. Similarly, of those who were not rearrested, black defendants were more likely to have been classified as high risk than white defendants. So white defendants who turned out to be risky were released and black defendants who turned out not to be risky were held. While no model will predict with perfect accuracy, ProPublica's conclusion was that the model is racially biased because the predictions were inaccurate more often for black defendants than white defendants. Therefore, the algorithm violated the second and third criteria described above, providing false positive and false negative rates that differed by race (Chouldechova & G'sell, 2017).

In its response to ProPublica, Northpointe did something different: it worked forward from the risk score instead of backward from the outcome. It found that people with similar risk scores, whether black or white, had similar chances of getting rearrested. In other words, they compared the predictive value of the score across racial groups and found that they were similar (Flores et al., 2016).

Thus, Northpointe makes a technically valid argument using a very different measurement, illustrated in Figure 2. They found that of those who were classified as high risk, the proportion who were not actually rearrested was roughly equivalent between both white and black populations. Similarly, they found that of those who were classified as low or medium risk, blacks and whites had a roughly equal chance of being rearrested. Northpointe concludes that these equivalent proportions "exhibit accuracy equity," or predictive fairness, which they argue should be the assessment metric for fairness in risk models. Northpointe is using the first criterion described above: the predictive value of the model across different groups. This does not

change the number of people in each group analyzed above. Instead, it rearranges them, comparing people who have the same risk score at the outset, rather than the same behavioral outcome.

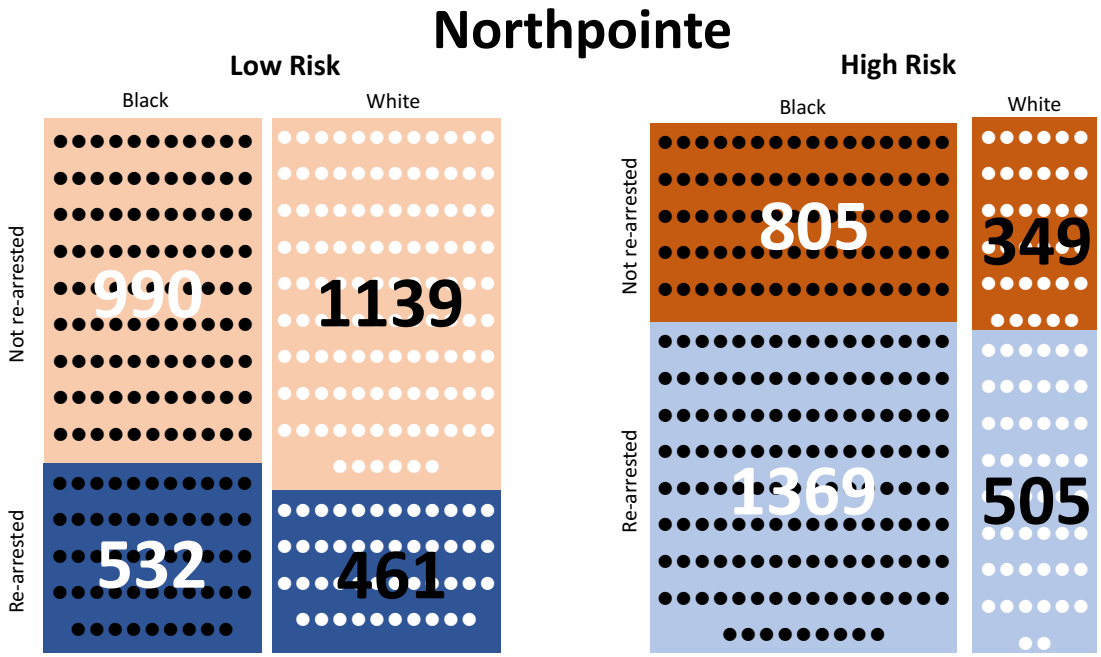

Figure 2. Northpointe's Analysis: categorize defendants by their risk score, then measure how many of those in each category reoffend. The share of black and white defendants with the same risk score who go on to be rearrested is similar and thus, the bars are similar heights.

To understand Northpointe's rebuttal, note that there is a substantial difference in the overall rate of rearrest for black and white defendants in this analysis: 51% of black defendants were rearrested versus 39% of white defendants. This difference means that models will predict that a greater proportion of black defendants will be rearrested than white defendants, since models assume that the future will be like the past. Because the model predicts that a greater fraction of black defendants than white defendants will be rearrested, a greater fraction of black

defendants could potentially be misclassified as probable rearrestees.[6] Recall, though, that in circumstances with big differences in rearrest across groups, it is impossible for a model to be fair on the first criterion as well as on the second and third criteria identified above (Chouldechova, 2017; Kleinberg et al., 2016). Thus, it is impossible for a model to be fair on both Northpointe's and ProPublica's terms: unless groups have similar recidivism rates, the model cannot have both equal predictive validity and equal rates of false positives.

Proxies for Race, Gender, and Class

Since the ProPublica and Northpointe debate, there have been many outspoken concerns over racial bias in risk assessment models. Nevertheless, proponents of these models sometimes argue that if a model does not include race as a variable, it is race neutral. However, there are two major problems with this claim. First, many of the variables used in these models act to some extent as proxies for race. This is especially true for criminal history, a variable which seems highly relevant on its face but is strongly influenced by race (Hannah-Moffat, 2013; Harcourt, 2015). We unpack this problem in significantly more detail in the following section.

Second, when a variable like race is excluded from a model, the estimates of the impact of other variables— that correlate with both the excluded variable and the outcome—will incorporate the effect of the missing variable on race. For example, the defendant's neighborhood or zip code might be included in the model, and many neighborhoods and zip codes are inhabited by people mostly of one race. Indeed, social scientists have had difficulty finding comparable black and white neighborhoods in order to study the consequences of

[6] People who are classified by the model as "high risk" but are not rearrested are called "false positives," and their share of the total number of people is called the "false positive rate."

poverty for child development (Perkins & Sampson, 2015), because neighborhood segregation is related to so many factors. In those cases, the defendant's zip code will act in the model as though it is partially a proxy for race. In a society structured by racism and segregation, many variables commonly included in models, from location to employment to prior police encounters, will be correlated with race. When the model is used to make predictions, those estimates of the effect of different variables that are correlated with race will be used to calculate risk assessment scores, thus incorporating information about race into the assessment.

This problem—that it is impossible to exclude race from the models—means that risk assessment instruments may not be able to overcome constitutional equal protection concerns. Even if the models do not include race or other "highest tier" protected classes, proxies for race - like zip code, income-level, education level and, perhaps most crucially, number of prior police encounters - are commonly used in this model. They will continue to include the effect of race unless models omit all variables that are both correlated with the outcome of interest and with race.

One methodological alternative is to directly consider protected variables such as race and gender in algorithms. However, this raises substantial issues for due process and equal protection. The COMPAS tool uses gender as a factor (*Wisconsin v. Eric Loomis*, 2016). Here, it is worth unpacking precisely *how* the COMPAS tool uses gender: excluding gender, women and men with the same score have negative outcomes at very different rates: "female defendants with a risk factor of 6 recidivated about as often as men with a factor of 4" (Drösser, 2017, para. 10). Thus, without considering gender as a factor, the COMPAS tool fails the first statistical fairness criterion described above, "calibration within groups" or "test fairness." A second alternative, described by Zliobaite (2017), develops prediction scores with protected variables

included, then applies the model for one group to all members. Thus, for example, researchers might use the predicted scores for white defendants regardless of the defendant's race. This approach does provide a clear way to address omitted variable bias, but these scores would still need to be evaluated using the incompatible criteria described above.

One of the few court cases to consider the constitutionality of risk assessment instruments, *Wisconsin v. Eric Loomis* (2016), engaged only the due process implications of including gender as a factor in the COMPAS instrument, setting aside equal protection concerns because the defendant did not raise them. The court found that considering gender had value to both the justice system and defendants more broadly (p. 35). It is by no means clear that this approach will meet with the approval of other courts. There is a long history of using actuarial prediction to make legal and judicial decisions. In *Craig v. Boren* (1976), the Supreme Court rejected an Oklahoma law that set different age standards by gender for buying alcohol, despite evidence that gender was a relevant predictor of alcohol abuse, arguing that gender had been consistently rejected as an appropriate factor for legal consideration.

When it comes to risk assessment instruments, there are factors beyond race and gender that could raise constitutional objections. As Starr explains, "the most widely used instruments incorporate much more detailed analyses of the defendants financial, housing, family, and employment history, current situation, and prospects" (Starr, 2015, p. 230). As we will see in unpacking the base layer of the model, this raises serious concerns about the extent to which individual defendants are being judged based on their group memberships.

Predictive Accuracy

One of the strongest arguments in favor of risk assessment is that data-driven risk assessment models offer increased predictive accuracy over professional, clinical, or

unstructured judgment in a variety of settings, including education, medicine, psychology, and criminal justice (Andrews et al., 2006; Grove et. al, 2000). Predictive accuracy means that the models correctly classify defendants' risk level: an important contributor to fair and efficient decision-making. After reviewing several studies to this effect, Desmarais, Johnson, and Singh (2016) conclude there is "overwhelming evidence" that risk assessment instruments, including COMPAS, result in superior predictive accuracy to human judgment (Brennan et al., 2009; Desmarais, Johnson, & Singh, 2016, p. 206).

However, recent studies of untrained human judgment showed comparable performance between the predictive accuracy of the COMPAS tool and predictions produced by untrained participants (Dressel & Farid, 2018). Despite reported differences in predictive accuracy between risk assessments and human judgments, several studies have found that most risk assessments on the market perform comparably in terms of AUC, a common measure of predictive accuracy. (Monahan & Skeem, 2016; Yang et al., 2010). There is no clear standard for the level of predictive accuracy needed to justify using models for high-stakes questions like liberty decisions (Yang et al., 2010).

Interpretive Fairness

While statisticians have focused on the aforementioned criteria for assessing the fairness of risk prediction scores, one often overlooked aspect of algorithmic fairness is the way that the risk scores are translated for judges. Judges rarely see the raw output of the model itself – the percentage chance that someone will commit another crime or fail to show up for trial. Rather, risk assessment tools often group defendants' raw risk score into ordered categories for easier interpretation (i.e. low risk, medium risk, high risk). This categorization, whether accomplished by human decision-making or technical processes, is itself an integral part of any risk assessment

model and can distort assessment of risk. Thus far, fairness assessments have not developed criteria for deciding whether models fairly translate the risk into ordinal categories.

There may be substantial interpretive issues with these categories, making it difficult to determine what a risk score actually means. One natural interpretation, which we call the intuitive interpretation, is that ordered categories are similar sizes and cover the full spectrum of approximate risk levels. For example, on a five-point scale, category one would mean a 0–20% risk of reoffending (or whatever outcome of interest), category two would mean a 21–40% risk, and so on and so forth. However, this is not the case for many, perhaps most, scales for which we have published data.

The Pretrial Risk Assessment Tool (PTRA) developed for the Administrative Office of the U.S. Court converts risk scores into a five-point risk scale (Lowenkamp & Whetzel, 2009). In Table 1, we compare the intuitive interpretation of the five-point scale to the actual likelihood of the outcome (in this case, rearrest, violent rearrest, failure to appear, and/or bail revocation). Only in the case of the lowest risk level is the actual probability of any negative outcome within the intuitive interval for that score. In the highest risk level, 65% of defendants have no negative pretrial outcomes. In other words, only 35% of defendants classified at the highest risk level failed to appear for trial or were rearrested before trial. The probabilities of failure to appear and rearrest for all risk levels, even the highest, are within the intuitive interval for the lowest risk level. Even more concerning, technical violations make up an increasing share of the negative outcomes as the reported risk level rises. In the lowest risk level, 33% of reported negative outcomes are technical violations; by the highest risk category, 42.8% are (Lowenkamp &

Whetzel, 2009).[7]

Further research on the PTRA confirms this mismatch between intuitive interpretations and actual risks. In a recent study on the PTRA, less than 20% of defendants in the highest risk category were rearrested for any crime, only 3.8% were arrested for a violent crime, and only 4.9% failed to appear in court (Austin, 2017). In some cases, analysts report these actual differences only deep in the appendix, treating the decision about how to convert predicted probabilities into categorical scores as a minor aspect of their work—despite the deeply unintuitive results it produces (VanNostrand & Keebler, 2009).

Table 1. *Intuitive vs Actual Interpretations: PTRA Design*

Risk Score	Intuitive Interpretation	Actual Violations	Actual (excluding technical violations)
Risk Level 1	0–20%	3%	2%
Risk Level 2	21–40%	10%	6%
Risk Level 3	41–60%	19%	10%
Risk Level 4	61–80%	29%	15%
Risk Level 5	81–100%	35%	20%

Note. Adapted from "The development of an actuarial risk assessment instrument for US Pretrial Services" by Christopher T. Lowenkamp and Jay Whetzel, 2009, *Federal Probation Journal, 73*(2), p.36.

While the example of the PTRA is striking, it by no means stands alone. In a study of the PSA-Court, the tool in widest use in state jurisdictions, the rate of rearrest on a violent charge for defendants classified as high-risk for violence was only 8.6% (Mayson, 2016). In Colorado's

[7] As we will see in the following layer, rearrest as an outcome measure is biased by racial bias in the criminal justice system; it also does not differentiate between, for example, drug use and violent crime.

Pretrial Assessment Tool, there are substantial gaps between the intuitive and the actual interpretations of risk categories (Pretrial Justice Institute, 2013). When the overall prevalence of the outcome is low, the difference between the actual versus intuitive interpretation of the risk score increases with each risk category. This poses risks to defendants. If, in the case of the PTRA study, judges and decision-makers observe a score of five and believe it means defendants are more likely than not to skip bail or be rearrested before trial, the categorization erroneously inflates the risk. In reality, a very small percentage of the defendants assessed in the PTRA study had negative pretrial outcomes—something that the categorization system obscures (Lowenkamp & Whetzel, 2009).

Policy Recommendations for the Model Fairness Layer

Balancing different statistical measures of fairness requires courts and policymakers to decide which type of fairness is most important: accurate prediction or equalizing false positives and false negatives across groups. These questions cannot be answered by statisticians. However, since these tradeoffs are inevitable, they should be made explicitly. Defendants and defense lawyers should be able to analyze the fairness—and the criteria used to measure fairness—for the models used to make liberty determinations about their cases. Policymakers should debate the value of different criteria for fairness as they choose which models to adopt.

In addition, risk assessment models should explicitly describe the process by which a predicted probability of failure to appear or a predicted probability of rearrest are translated into risk scores. Moreover, we need substantially more research on the ways that judicial decision-makers interpret risk scores. Do they use the intuitive interpretation described above, do they norm to the level of failure to appear or rearrest they observe in their own courtrooms, or do they use some other process such as blame avoidance to translate scores into policy outcomes?

Researchers can help with this, but they need access to the categorization scheme and predicted probabilities in the model itself to make effective assessments. Furthermore, they need access to the data that was used to create the model to assess the concerns raised in the middle layer, to which we now turn our attention.

Middle Layer: Data Quality

Fundamental Problems

The second layer in the debate about fairness involves whether or not the data a risk assessment models is based on is biased. For both ProPublica and Northpointe, the measure of "risk" was whether the defendant was rearrested while on probation (Angwin et al., 2016; Dieterich et al., 2016). This use of arrest as a measure of criminality fundamentally assumes that people who do the same things are arrested at the same rates.

However, there is plenty of evidence that people of color, especially black people, are more likely to be arrested than whites for the exact same behavior. Black Americans are disproportionately likely to be stopped and searched by police, whether they are driving or walking (Epp, Maynard-Moody, & Haider-Markel, 2014; Gelman, Fagan, & Kiss, 2007; Goel, Rao, & Shroff, 2016; Harcourt, 2015). White and black Americans use marijuana and other drugs at similar rates, but black Americans are much more likely to be arrested for drug possession (Edwards, Bunting, & Garcia, 2013; Epp et al., 2014; Goel et al., 2016; Simoiu, Corbett-Davies, & Goel, 2016). This is a problem for the criminal justice system, but it is also a problem with criminal justice data.

Imagine two identical young men. Greg is white and Jamal is black.[8] They live on the same block, they take their children to the same neighborhood schools, they smoke marijuana with identical frequency, and they drive identical cars at identical speeds. Greg and Jamal commit the same crimes and have done so their entire lives. They should look identical to a risk assessment model. What the model knows about Greg and Jamal's criminal histories, though, comes primarily from their arrest records. Arrests are the result of the combination of individual behavior and police decisions. If Jamal is more likely to get noticed, followed, stopped, searched and arrested by police—and the evidence suggests he is (Epp et al., 2014; Harcourt, 2015; Gelman, et al., 2007; Goel et al., 2016) —their identical behavior will translate into radically different data, with more arrests for Jamal. Thus, the model's predictions will score Jamal as riskier, even though he and Greg have lived identical lives. This disparity would be exacerbated if they lived in different communities with different levels of policing; since communities of color are often under greater surveillance than white communities (Goffman, 2014).

In fact, the model would, from a statistical perspective, be completely correct in scoring Jamal as having a higher risk of rearrest. The problem is that being arrested is a racially unfair measure of whether a person is truly dangerous to the community. By using arrest as a measure of criminality, the model bakes in the fact that Jamal is more likely to be arrested because he is a person of color (Edwards et al., 2013; Epp et al., 2014; Gelman, et al., 2007; Goel et al., 2016; ; Spencer, Charbonneau, & Glaser, 2016). Racial bias in arrests leads to racial bias in risk scores.

The same is true for other measures drawn from the criminal justice system. Defendants

[8] These names are taken from a well-known study of the effect of racism on employment opportunity (Lavergne & Mullainathan, 2004).

of color, especially black and Latino men, are treated more harshly throughout the criminal justice system (Kutateladze, Andiloro, Johnson, & Spohn, 2014), disproportionately prosecuted (Beckett, 2012; Western, 2007), less likely to be offered pretrial diversion, counseling, and other supportive programming (Schlesinger, 2013), and sentenced to longer terms (Stolzenberg, D'Alessio, & Eitle, 2013). Criminal history, from arrest to conviction to sentencing to recidivism, is a measure of criminal justice practices that systematically target race-class subjugated communities (Soss & Weaver, 2017), not a measure of individual behavior.

The magnitude and pattern of the bias in the data cannot be measured directly with the techniques used by ProPublica, Northpointe, or any of the others arguing about these models, including us. Even if we accept Northpointe's argument that their risk assessment models make predictions that are equally likely to be right (or wrong) for black and white defendants, the models are built from data that makes people of color look riskier than whites, so the predictions will necessarily be biased. To make matters worse, the predictions will seem correct. The model is trained on data generated by past police bias, and we are asking the model to predict events that are dependent on future police bias. There is a perfect circularity to the model building and assessment, masked by the technical complexity of the discussion.

The model is not predicting solely individual behavior, but an event influenced by police decision-making. Once Greg and Jamal are released on probation, where they continue their identical lives, Jamal is still more likely to be arrested than Greg. It looks like the model correctly predicts that Jamal is riskier, but both the prediction and the outcome are the result of racially biased law enforcement.

This problem is not easily solved. Using arrests and other criminal justice data as an unbiased source of information on individual behavior would require us first to build a racially

unbiased criminal justice system. This is implausible on any reasonable timeframe. More independent data efforts could allow us to identify the magnitude of the problem but would not eliminate policing-induced bias from individual-level tools. Ultimately, the problem of bias in the data is a serious threat to the entire endeavor of data-driven risk assessment. When both the data used to produce the risk assessment instrument and the data used to evaluate it come from the criminal justice system, quantitative risk assessments merely launder that bias. In other words, the legitimating process of quantitative assessment converts unequal data-generating processes into apparently objective data without removing the fundamental problems (Goddard & Myers, 2017; Ward, 2015).

Solvable Problems

Aside from these fundamental problems, there are data issues with prediction tools that are more amenable to technical solutions. Most centrally, there are often problems with predictions in heterogeneous populations. Recall that the model used to evaluate Emily (as well as Greg and Jamal) predicts Emily's risk based on data from other people: a sample. What if Emily (or Greg or Jamal) is simply very different from other people in the data? For example, crime patterns in the United States have changed dramatically over the last several decades and it is not clear that the factors that predict recidivism have remained constant over that period. Similarly, we might wonder whether predictions based on data from one jurisdiction generalize to the country as a whole.

One of the simplest ways to improve prediction is to make the sample larger and more diverse, so that predictions for individuals can be based on data about "similar" defendants. However, "similarity" is challenging to judge, and requires collecting additional data. Researchers with the Laura and John Arnold Foundation report that their tool's predictions,

while based on a particular set of data, are similarly accurate across multiple jurisdictions, though they have not released information that allows other researchers to validate these reports (Laura and John Arnold Foundation, 2018). Their tool is also based on a large data set: over 1.5 million cases from 300 jurisdictions (Laura and John Arnold Foundation, 2018). In contrast, some of the validation studies for Northpointe's COMPAS tool are based on data sets as small as 2,328. The number of people with a specific combination of variables may be extremely small, or indeed zero. On a population that small, the tool will be difficult to validate both for small populations and for any analysis considering multiple axes of marginalization.

However, expanding the sample size comes with analytical problems relevant to the model fairness layer. For example, with a large enough sample size, models may show statistically significant differences between groups even when the substantive differences are not large enough to seem important. For example, in the PTRA model described above, it is unlikely that the difference in frequency of violations (rearrest or failure to appear) between risk level four (29%) and risk level five (35%) occurred by chance. With over half a million observations, the model has plenty of power to identify small non-random differences (Lowenkamp & Whetzel, 2009). Yet it is not clear that judges should see a substantively meaningful difference between a 29% and a 35% chance of a problem pretrial outcome.

Some data fairness problems in this middle layer are solvable. The reality, though, is that all of our data will be biased in ways that make defendants targeted by the carceral state look more dangerous both to the initial risk assessment instrument and to those evaluating its fairness. The circular reasoning of predictive modeling is not new (Blackmon, 2008), and it will continue to pose tremendous obstacles to developing a fair version of data-driven risk assessment.

Base Layer: Fundamental Conceptual Problems with the Fairness of Data-Driven Decisions

These questions of algorithmic fairness and data quality affect all types of algorithms, from those used to select who sees which advertisements on Facebook to those used in the criminal justice system. However, when predictive models are applied to core state decisions—especially liberty decisions like pretrial detention or sentencing—they raise an even deeper set of constitutional, legal, and conceptual concerns about fairness: a third layer of this debate. Even if a risk assessment model was statistically fair and based on unbiased data, there is still a fundamental problem: it evaluate a defendant's risk using data about *other people*. The risk assessment instrument uses information about a group of people which does not include the defendant, and provides a score based on *their* behavior. Is it fair to evaluate Emily or Greg or Jamal's risk based on the behavior of a group they belong to—however narrowly tailored that demographic and behavioral group might be? For Emily, Greg, and Jamal, their risk scores are the result of other people's behavior, not theirs.

As a constitutional matter, defendants are entitled to have their sentence based on what they personally did, rather than based on what people who share their social, demographic or geographic group affiliations did. While statistical models that use group-based averages may produce better predictions than decisions made with inadequate information or inconsistent criteria, "the Supreme Court has held that this defense of gender and race discrimination offends a core value embodied by the Equal Protection Clause: people have a right to be treated as individuals"(Starr, 2014, p.16).

Data-driven risk assessment *inevitably* incorporates information about race, gender, and other protected categories. The specific factors most often considered exacerbate this problem.

As Harcourt (2015) baldly puts it, "The fact is, risk today has collapsed into prior criminal history, and prior criminal history has become a proxy for race" (p. 237). The inclusion of socioeconomic variables worsens this problem, as risk assessments then treat race and class inequality as a personal quality that makes a defendant riskier (Goddard & Myers, 2017). Risk assessment instruments directly include measures of housing stability, employment history, debt, and numerous other factors closely correlated with race and class (van Eijk, 2017). These tools thus directly encourage judges to treat defendants "more harshly because they are poor or uneducated, or more lightly because they are wealthy and educated" (Starr, 2014, p. 804).. Moreover, they shift attention and resources from structural factors – like addressing racial inequality and poverty – to individual measures of those structural factors (Goddard & Myers, 2017; van Eijk, 2017).

However, even a determined effort to exclude proxies for race, class, or other marginalized categories is not likely to be successful. As described above, omitting race from the set of variables in the original data set does not mean race is not included in the analysis: it merely induces remaining variables that are correlated with both race and the outcome variable to behave as if they are in part proxies for race. Thus, risk assessment instruments may not be able to overcome constitutional equal protection concerns.

Regardless of the intent or apparent neutrality of the instruments, government agencies are also subject to the scrutiny of state and city constitutional and statutory guarantees, like those enacted in New York City, to assure they do not have a disparate impact on a protected class (N.Y.C. Admin. Code §14-151). Some of the concerns raised in the discussion of model fairness, and in the discussion of bias in the criminal justice system, mean that risk assessment models are likely to affect black and white defendants differently. If these risk assessments are racially

discriminatory, because of a flaw in the model (a methodological critique) or because the data are inherently biased, using them could also result in a disparate impact.

While very few of these issues have been litigated, they are important to understanding fundamental questions about whether data-driven risk assessment can be fair. The Wisconsin v. Eric Loomis case was the closest any court has come to addressing this argument when the defendant challenged the risk assessments recommendation of prison instead of probation. Because the court found that the sentencing judge had "explained that its consideration of the COMPAS risk scores was supported by other independent factors" and therefore, "its use was not determinative" the court decision did not prohibit the consideration of the risk assessment in sentencing decisions, or require COMPAS to disclose the underlying data or methodology of the risk assessment (*Wisconsin v. Eric Loomis*, 2016, p. 5). Loomis did, however, recognize the constitutional danger in relying solely on risk assessments in the decision-making process, prohibiting the use of risk scores as the decisive factor in liberty decisions (*Wisconsin v. Eric Loomis*, 2016).

Additionally, Loomis cautioned courts using risk assessments that they are only able to identify a group of high risk offenders and not a particular high risk individual, that "an offender who is young, unemployed, has an early age at first arrest and history of supervision failure, will score medium or high on the COMPAS Violence Risk Scale even though the offender never had a violent offense" (*Wisconsin v. Eric Loomis*, 2016, p. 29). Loomis resolves the problem that risk assessment models provide information about groups rather than individuals by requiring judges to consider other factors in sentencing or pretrial release decisions (*Wisconsin v. Eric Loomis*, 2016, p. 31), though, as we argue below, this finding raises concerns about how human judgment and risk assessments interact.

Secret Models, Secret Data, and Due Process

Additionally, the Wisconsin Supreme Court's Loomis decision failed to remedy the fact that predictive models in criminal justice are typically secret. At a minimum, the data used to develop the model are secret, though sometimes, as with the Arnold Foundation's Public Safety Assessment, the methods used to calculate the score are made public (Laura and John Arnold Foundation, n.d.). Loomis argued "that because COMPAS does not disclose this information, he has been denied information which the circuit court considered at sentencing" (*Wisconsin v. Eric Loomis*, 2016, p. 21). The court held that, because the defendant and the court saw the same information, the defendant was not entitled to information about how the scores were calculated or evaluated (p. 23).

This is a deeply unsatisfying holding for anyone concerned about the discriminatory potential of risk assessment instruments. By allowing Northpointe's trade secrets claim to stand, the Loomis court prevented judges, defendants, and researchers from vetting the algorithms and evaluating the fairness of both the top and middle layers. By keeping defendants from challenging their risk scores, these protections "signal . . . that the government values trade secrets holders as a group more than those directly affected by criminal justice outcomes" (Wexler, 2017, p. 5). They also suppress information to judges themselves by preventing researchers and defendants from providing fair and thorough evaluations of the risk assessment instruments.

For example, ProPublica's analysis of the COMPAS tool used the scores for over 10,000 defendants in Broward County, FL, over a two-year period (Angwin et al., 2016). ProPublica could not examine the fairness of the COMPAS score unless it was used by a public entity with available records. To generate that analysis, over 10,000 defendants needed to actually be

assessed and sentenced using the COMPAS tool: these individuals formed an important pool for analysis, but the analysis would not have been possible without influencing their liberty decisions. This places defendants in the role of research subjects except with none of the protections institutional review boards insist upon. If a scoring tool is found to be biased based on its actual application, that finding comes too late for thousands of defendants whose liberty decisions were affected by the biased tool.

Courts need to ensure that researchers, defendants, and judges have access to information that allows them to understand the problems in specific risk assessment tools, since those problems will be quite different depending on the details of the data, methodology, and conversion to risk scores. Of course, determining what specific information is required for an adequate assessment is somewhat more complex. Companies, non-profits, and researchers creating risk assessment instruments have legitimate concerns about the privacy of the subjects observed in the data they use to generate the model. Releasing that data publicly could lead to embarrassing, stressful, or damaging disclosures to which individuals have not consented.

Kroll et al. (2016) provide a system for establishing transparent, replicable algorithms. In their system, designers of algorithms disclose how decisions are made. They provide an example of the Diversity Visa Lottery, and suggest that designers should specify both the *inputs* and the *mechanism* by which those inputs are used to generate a score (Kroll et al., 2016). This is uncommon but not unheard of in risk assessment: the Arnold Foundation's Public Safety Assessment, one of the most widely used tools, discloses the scoring formula on its website (Laura and John Arnold Foundation, 2018).

Still, these disclosures are not sufficient to evaluate the fairness of algorithms in even the top layer. For that, researchers need access to the judgments about specific individuals and the

outcomes for those individuals after the liberty decisions are made. Disclosing this sensitive data publicly may be a problem but anonymized versions can be released under a protective order to defendants (Wexler, 2017), or to researchers operating under the supervision of an institutional review board.

Addressing the middle layer is even more challenging, since neither developers nor researchers have a reliable source for data on individuals that is not compromised by biases in the implementation of the criminal justice system. One partial solution is for stakeholders interested in fair algorithms to fund the collection of alternative data sources, which can be used to assess the consequences of bias in criminal justice data. This does not fully resolve the problems raised in the middle layer, but it would be a first step towards examining the consequences of the errors currently baked into risk assessment models.

Transparency in *both* risk scoring and training data is a necessity for researchers to be able to vet risk assessment instruments. Instruments vary widely, and each one needs to be examined for fairness individually. Neither the top nor the middle layer of fairness considerations can be examined without information about the way the algorithm is constructed, the data on which it is based, and its consequences for different race, gender, and class groups.

Human Judgment Is Also Biased

Ironically, much of the advocacy for risk assessments stems around the need for transparency in criminal justice decision-making. In practice, we often do not know the reasoning behind a judge's decision. Research has indicated that there are three focal concerns judges typically have when making sentencing decisions: the defendant's blameworthiness, potential dangerousness (i.e. recidivism risk), and practical/organizational constraints (Steffensmeier, Ulmer, & Kramer, 1998, p. 788). However, judges often do not have sufficient

information to make informed predictions . They may also fall victim to many of the aforementioned problems with quantitative analysis.

Like algorithms, human judgment is also based on shortcuts that take into consideration social, demographic, geographic, and behavioral patterns (Banks et al., 2006; Eberhardt et al., 2006; Kang et al., 2011; Levinson, 2007; Levinson et al., 2010; Richardson & Goff, 2012). Steffensmeier et al. (1998) posited that judges often develop a "perceptual shorthand" that incorporates their own perceptions and stereotypes – often based on protected characteristics - into their decisions. Humans also have difficulty making fair decisions when there are multiple considerations to weigh (top layer); humans are exposed to biased data, which may influence their perceptions of risk (middle layer); and humans apply their experiences of groups to their decisions about individuals (base layer). Indeed, there is substantial evidence that defendants of color are disadvantaged in pretrial and sentencing decisions made without reference to risk assessment models (Demuth, 2003; Kutateladze et al., 2014; Menefee, 2018; Steffensmeier & Demuth, 2000; Steffensmeier et al., 1998; Woolredge et al., 2015).

In this context, risk assessment can offer increased transparency and standardization in pre-trial recommendations, because the relevant factors under consideration are clearly enumerated and consistent (Summers & Willis, 2010; Lowenkamp, 2008; Lowenkamp & Whetzel, 2009). Cooprider (2009) reports similar advantages in that the standardization of decision-making processes "minimiz[es] arbitrariness, individual bias, and systemic disparity" (p. 13). Additionally, risk assessment supports an "operational definition of justice" in that individuals with similar backgrounds and charges would receive similar bond amounts (Cooprider, 2009, p. 13).

Furthermore, human judgment is biased, many proponents of quantitative risk assessment

tools claim their biased outcomes are likely to be an improvement to the current system (Corbett-Davies et al., 2016). If judges are making misinformed decisions, or decisions based on their own implicit biases, then a statistical analysis may be an improvement to the current status quo. However, there is currently imperfect evidence that the biased results of quantitative models will be superior to the biased results of human judgment. While Desmarais et al. (2016) conclude that quantitative models are much more accurate than human judgment, a preliminary study comparing untrained human prediction with COMPAS found human judgment to be very slightly more accurate (67% correct, compared to 65% for COMPAS), with no difference in racial bias by either metric for algorithmic fairness (Dressel & Farid, 2018).[9] Furthermore, using quantitative risk assessment models does not eliminate the role of human judgment. Rather, the Loomis decision asks judges to consider the score as one aspect of an individualized decision—effectively layering the problems of human judgment on to the technical problems of algorithmic fairness, data quality, and translation from probability to risk score (*Wisconsin v. Eric Loomis*, 2016).

Indeed, the presence of risk scores may change how judges make decisions. Rather than substituting the risk assessment instrument for their existing judgment about risk, it may shift judicial attention from deservingness or other factors to risk, arguably making demographic factors more, rather than, less salient. Starr (2014) explains that judges understand the challenges of predicting recidivism, and may therefore set recidivism aside in their decision-making.

[9] Note also that the human judgment, like the model, is evaluated using the same biased data we critique in our discussion of the middle layer.

Offered a risk assessment instrument, that same judge may place more weight on recidivism risk, believing that they now have a means to evaluate it well. In an experiment in which criminal law students were shown cases with and without risk scores, Starr finds that including predicted risk scores increased the weight participants gave to recidivism risk as opposed to other sentencing considerations (Starr, 2014).

Proponents of risk assessment tools argue that they will increase judicial willingness to allow pretrial release. The recent work of Kleinberg et al. (2016) shows through policy simulations that if risk assessment were implemented perfectly such that all judges followed its recommendations, more defendants could be released while maintaining similar levels of measured recidivism and failure to appear and decreasing racial disparities. Other scholars have argued that risk assessment may play a central role in unwinding mass incarceration (Monahan & Skeem, 2016). In practice, when risk assessment is introduced, judges may still exercise discretion, and evidence about the efficacy of risk assessment in practice is limited and mixed.

The logic connecting risk assessment to decreased pre-trial detention is simple. Judges often worry about the risk of releasing someone (before trial or via a shorter sentence) who goes on to commit a serious or violent crime, both because they care about protecting their communities and because they worry about public backlash. A risk assessment tool offers a neutral alternative: if the judge follows the outcome suggested by the risk score, they can engage in what political scientists call "blame avoidance" (Weaver, 1986). The inverse is also true: a judge concerned with avoiding blame for a problem decision will be unwilling to release someone with a high risk score, even if, as described in the PTRA study above, the absolute risk of both failure to appear and violent recidivism is extremely low (Lowenkamp & Whetzel, 2009). This is likely to be especially appealing to "'elected' judges and prosecutors who must

defend their decisions to an electorate concerned with security (Hannah-Moffat, 2013).

Thus far, some evaluations of risk assessment have in fact shown reductions in detention. A randomized control trial of the Harlem Parole Re-entry Court, which combined risk assessment with additional services, found several improved outcomes, including lower rates of re-arrest and reconviction (Ayoub & Pooler, 2006). Some evaluations of the Arnold Foundation's Public Safety Assessment found that judges do in fact release more defendants when offered risk assessment tools (Laura and John Arnold Foundation, 2016; Schuppe, 2017). In Kentucky, though, after an initial drop in pre-trial detention, judges eventually reverted old rates of pre-trial detention experienced prior to the introduction of the new risk assessment tool (Stevenson, 2018). Yet another implementation site, Lucas County, OH, found an increase in pre-trial detention (*Jones v. Wittenberg,* 2017).

Recent incidents raise questions about whether the logic of blame avoidance will hold as these tools become more popular. In San Francisco, a person with a low risk score was released and, five days later, committed a murder in the course of a robbery; the result was an outcry about the release decision (Westervelt, 2017). The situation is similar to that of body-worn cameras among police: they appeared to radically change police behavior in preliminary tests (White, 2014). Later tests, conducted once cameras had become commonplace and been assimilated to existing political dynamics, found no effect (Ariel et al., 2016; Yokum, Ravishankar, & Coppock, 2017). The value of risk assessment tools in increasing the frequency of release for low-risk defendants depends almost entirely on how they interact with local political dynamics and enable the logic of blame avoidance.

There are other mechanisms to enforce transparency and reduce detention that do not require homogenizing decisions with risk assessment instruments. For example, regulations

could be implemented to require on the record, formulaic rationalization for pre-trial decisions. These records would illuminate the focal concerns judges are considering when making their decisions and if desired, could be formulated to prompt them to consider, or at least speak to, factors they may have otherwise ignored. Additionally, such rationalizations would produce data that could be analyzed to uncover bias, even if obscured in their rationalizations, by contrasting their decisions for otherwise similar members of different protected classes. Furthermore, it would provide a feedback loop for judges to identify implicit biases in their decisions, measure their accuracy of their predictions, and compare the severity of their decisions to other judges. Unlike risk assessment instruments, this type of data collection would increase judicial accountability and transparency to the public.[10]

Conclusion

This article proposes a new framework for thinking about problems with data-driven risk assessment tools. Existing literature on the problems of algorithmic fairness typically focuses on one of the layers: computer scientists, statisticians, and quantitative social scientists develop mechanisms for addressing the problems of the top layer (Chouldechova, 2017; Corbett-Davies et al., 2017; Feldman et al., 2015; Kleinberg et al., 2016), while legal scholars focus on the constitutional concerns related to the base layer (Hannah-Moffat, 2013; Harcourt, 2015; Starr, 2014).

[10] A recent study was able to compare judges' bail-setting practices and illuminated which ones over or under used bail (Barry-Jester, 2018). If demographic data on the defendants were available, similar cases could be compared to identify if judges are making biased decisions based on a protected characteristic.

In considering the fairness of algorithms, courts and decision-makers should not be distracted by the lack of intentional bias, or the promise of computer objectivity: "the resulting discrimination is almost always an unintentional emergent property of the algorithm's use rather than a conscious choice by its programmers," but this makes it "unusually hard to identify the source of the problem or to explain it to a court" (Barocas & Selbst, 2016, p. 671). Instead, courts and decision-makers should demand full evaluations of all three layers of bias.

In every layer, different ideas about fairness make the discussion harder to untangle. As Northpointe argues, their COMPAS model is "fair" in the sense that considering the underlying differences in arrest patterns, the model is about equally accurate at predicting rearrest for white and black defendants (Dieterich et al., 2016). If we think rearrest is a good measure of dangerousness, if we think the criminal justice system is equally fair for white and black people, then Northpointe's model and the resulting risk scores are also fair. In contrast, by focusing on the disparate outcomes for white and black defendants scored by the model, ProPublica implicitly proposed a different notion of fairness in the top layer.

The middle layer brings in questions about the larger criminal justice system that produced the data. We cannot assume that predictive fairness among risk groups makes Northpointe's model "fair" in the usual senses we mean when talking about justice. The problem of biased policing data in the middle layer is much bigger than one vendor's model. The data used to build these models carries the bias with it and the models then learns it. This is true for all criminal justice uses of data, but also for other algorithms, like the ones that target ads, hire employees, and offer credit.

Underneath these arguments about statistical fairness and biased data lies the fundamental conceptual problem of the base layer: is it fair to alter the life chances and liberty

outcomes of individuals because of their demographic, geographic, and social characteristics? Models must use data about other people to predict risk. This is particularly concerning in the criminal justice system, where racial inequalities are both dramatic and highly consequential (Lerman & Weaver, 2014; Pettit, 2012; Soss & Weaver, 2017; Western, 2007).

Resolving these problems is challenging, and offering comprehensive solutions is beyond the scope of this paper. Policymakers should evaluate risk assessment tools based on all three layers: algorithmic fairness, data bias, and the inherent justice of using group-based decision-making. However, many potential solutions to criminal justice problems sidestep data evaluation. In considering solutions to bail reform, governments might adopt the PSA or a similar tool. They might also eliminate money bail entirely, limit the types of offenses bail gets set on, limit the amount of bail that can be set for certain classes of offenses, or provide resources to help defendants return to court (childcare, transportation, etc.). These and other solutions sidestep the issue of fixing risk assessment, while engaging the fundamental problems risk assessment is intended to solve.

At every layer of analysis, it is clear that statistical and computer reasoning can clarify what is at stake but cannot decide the correct path. The process of constructing these models requires human judgment about what fairness means, in mathematical terms, and when it is morally acceptable to judge people based on the behavior of others. Judges, policymakers, and politicians like to be able to point to numbers to justify their decisions. But even if the risk scores were unbiased (which they are not) the numbers don't speak for themselves. We have to use human insight and human judgment to decide what they mean and when we should use them. In doing so, policymakers and judges need to consider all three layers of bias, and develop legal frameworks that promote transparency, accurate measurement, and just decision-making.

References

Andrews, D. A., Bonta, J., & Wormith, J. S. (2006). The recent past and near future of risk and/or need assessment. *Crime & Delinquency*, *52*(1), 7-27.

Andrews, D. A., Bonta, J., & Wormith, J. S. (2011). The risk-need-responsivity (RNR) model: Does adding the good lives model contribute to effective crime prevention?. *Criminal Justice and Behavior*, *38*(7), 735-755.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). "Machine bias." *ProPublica*. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Ariel, B., Sutherland, A., Henstock, D., Young, J., Drover, P., Sykes, J.,…Henderson, R. (2016). Wearing body cameras increases assaults against officers and does not reduce police use of force: Results from a global multi-site experiment. *European Journal of Criminology*, *13*(6), 744–755.

Ayoub, L. H., & Pooler, T. (2015). *Coming Home to Harlem: A Randomized Controlled Trial of the Harlem Parole Reentry Court*. New York: Center for Court Innovation.

Banks, R. R., Eberhardt, J. L., & Ross, L. (2006). Discrimination and implicit bias in a racially unequal society. *California Law Review, 94*(4), 1169–1190.

Barocas, S., & Selbst, A. (2016). Big data's disparate impact. *California Law Review*, *104*(671), 721.

Barry-Jester, A. M. (2018, June 19). You've been arrested. Will you get bail? Can you pay it? It may all depend on your judge. *Five Thirty Eight*. Retrieved from https://fivethirtyeight.com/features/youve-been-arrested-will-you-get-bail-can-you-pay-it-it-may-all-depend-on-your-judge/

Beckett, K. (2012). Race, drugs, and law enforcement: Toward equitable policing. *Criminology & Public Policy*, *11*(4), 641-653.

Blackmon, D. A. (2008). *Slavery by another name: the re-enslavement of Black people in America from the Civil War to World War II*. New York: Doubleday.

Bonta, J., & Andrews, D. A. (2007). Risk-need-responsivity model for offender assessment and rehabilitation. *Rehabilitation*, *6*(1), 1-22.

Brennan, T., Dieterich, W., & Ehret, B. (2009). Evaluating the predictive validity of the COMPAS risk and needs assessment system. *Criminal Justice and Behavior*, *36*(1), 21-40.

Children's Research Center. (2008). The structured decision making model: An evidence-based approach to human services (pp. 1–27). *National Council on Crime and Delinquency.* Retrieved from http://www.nccdglobal.org/sites/default/files/publication_pdf/2008_sdm_book.pdf

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, *5*(2), 153–163. https://doi.org/10.1089/big.2016.0047

Chouldechova, A., & G'Sell, M. (2017). Fairer and more accurate, but for whom? ArXiv:1707.00046 [Cs, Stat]. Retrieved from http://arxiv.org/abs/1707.00046

Cooprider, K. (2009). Pretrial risk assessment and case classification: A case study. *Fed. Probation*, *73*, 12.

Corbett-Davies, S., Pierson, E., Feller, A., & Goel, S. (2016, October 17). A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear. *Washington Post*. Retrieved from https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-

be-racist-our-analysis-is-more-cautious-than-propublicas/?utm_term=.ee8de35f2ea7

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness (pp. 797–806). ACM Press. https://doi.org/10.1145/3097983.3098095

Craig v. Boren, 429 U.S. 190 (1976).

DeMichele, Matthew, Peter Baumgartner, Michael Wenger, Kelle Barrick, Megan Comfort, and Shilpi Misra. "The Public Safety Assessment: A Re-Validation and Assessment of Predictive Utility and Differential Prediction by Race and Gender in Kentucky." (2018).

Demuth, S. (2003). Racial and ethnic differences in pretrial release decisions and outcomes: A comparison of Hispanic, Black, and White felony arrestees. *Criminology*, *41*(3), 873-908.

Desmarais, S. L., Johnson, K. L., & Singh, J. P. (2016). Performance of recidivism risk assessment instruments in US correctional settings. *Psychological Services*, *13*(3), 206.

Dieterich, W., Mendoza, C., & Brennan, T. (2016). COMPAS risk scales: Demonstrating accuracy equity and predictive parity. *Northpoint Inc*. Retrieved from https://assets.documentcloud.org/documents/2998391/ProPublica-Commentary-Final-070616.pdf

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. Science *Advances, 4*(1). doi:10.1126/sciadv.aao5580

Drösser, C. (2017, December 22). In order not to discriminate, we might have to discriminate. *Simons Institute for the Theory of Computing*. Retrieved from https://simons.berkeley.edu/news/algorithms-discrimination

Eberhardt, J. L., Davies, P. G., Purdie-Vaughns, V. J., & Johnson, S. L. (2006). Looking deathworthy: Perceived stereotypicality of black defendants predicts capital sentencing

outcomes. *Psychological Science*, *17*(5), 383–386.

Edwards, E., Bunting, W., & Garcia, L. (2013). *The war on marijuana in black and white*. (Technical report). New York, NY: American Civil Liberties Union.

Epp, C. R., Maynard-Moody, S., & Haider-Markel, D.P. (2014). *Pulled over: How police stops define race and citizenship*. Chicago: University of Chicago Press.

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

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

Freitag, R., and K. Park. "The Structured Decision Making Model: an evidenced-based approach to human services." *Madison, WI: Children's Research Center* (2008).

Gelman, A., Fagan, J., & Kiss, A. (2007). An analysis of the New York City police department's "stop-and-frisk" policy in the context of claims of racial bias. *Journal of the American Statistical Association*, *102*(479), 813-823.

Goel, S., Rao, J. M., & Shroff, R. (2016). Precinct or prejudice? Understanding racial disparities in New York City's stop-and-frisk policy. *Annals of Applied Statistics*, *10*(1), 365– 394.

Goddard, T., & Myers, R. R. (2017). Against evidence-based oppression: Marginalized youth and the politics of risk-based assessment and intervention. *Theoretical Criminology*, *21*(2), 151-167.

Goffman, Alice (2014). *On the run: Fugitive life in an American city.* Chicago: University of Chicago Press.

Gouldin, L. (2016). Disentangling flight risk from dangerousness. *BYU Law Review*, *2016*(3), 837–898.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: a meta-analysis. *Psychological assessment*, *12*(1), 19.

Hannah-Moffat, K. (2013). Actuarial sentencing: An unsettled proposition. *Justice Quarterly*, *30*(2), 270–296.

Harcourt, B. E. (2015). Risk as a proxy for race: The dangers of risk assessment. *Federal Sentencing Reporter, 27*(4), 237–243.

Jones v. Wittenberg, 73 F.R.D. 82 United States District Court, N.D. Ohio, Western Division. Document #105, "Notice of Filing Copy of Presentation Assessing Impact of Public Safety Assessment." Filed January 9, 2017. Retrieved August 2018 from https://thecrimereport.org/wp-content/uploads/2017/08/Lucas-County-court-filing.pdf

Wisconsin v. Eric Loomis, 881 N.W.2d 749 (Wis. 2016).

Kang, J., Bennett, M., Carbado, D., & Casey, P. (2011). Implicit bias in the courtroom. *UCLA Law Review*, 59, 1124.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2017). Human decisions and machine predictions. *The quarterly journal of economics*, *133*(1), 237-293.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Kroll, J. A., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2016). Accountable algorithms. *U. Pa. L. Rev.*, 165, 633.

Kutateladze, B. L., Andiloro, N. R., Johnson, B. D., & Spohn, C. C. (2014). Cumulative disadvantage: Examining racial and ethnic disparity in prosecution and

sentencing. *Criminology*, *52*(3), 514-551.

Lavergne, M. & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *The American Economic Review*, *94*(4), 991–1013.

Laura and John Arnold Foundation. "New data: Pretrial risk assessment tool works to reduce crime, increase court appearances." Retrieved August 2018 from http://www.arnoldfoundation.org/new-data-pretrial-risk-assessment-tool-works-reduce-crime-increase-court-appearances/

Laura and John Arnold Foundation. "Public Safety Assessment." Retrieved August 2018 from http://www.arnoldfoundation.org/initiative/criminal-justice/crime-prevention/public-safety-assessment/

Lerman, A. E., & Weaver, V. M. (2014). *Arresting citizenship: The democratic consequences of American crime control*. Chicago: University of Chicago Press.

Levinson, J. D. (2007). Forgotten racial equality: Implicit bias, decisionmaking, and misremembering. *Duke Law Journal*, 57, 345–424.

Levinson, J. D, Cai, H., & Young, D. (2010). Guilty by implicit racial bias: The guilty/not guilty implicit association test. *Ohio St. J. Crim. L.*, *8*, 187.

Lowenkamp, C. T., & Whetzel, J. (2009). The development of an actuarial risk assessment instrument for US Pretrial Services. *Fed. Probation*, *73*, 33.

Lowenkamp, C. T., Lemke, R., & Latessa, E. (2008). The Development and Validation of a Pretrial Screening Tool. *Fed. Probation*, *72*, 2.

Mayson, S. G. (2017). Dangerous defendants. *Yale LJ*, *127*, 490.

Menefee, M. R. (2018). The role of bail and pretrial detention in the reproduction of racial

inequalities. *Sociology Compass*, *12*(5), e12576.

Monahan, J., & Skeem, J. L. (2016). Risk assessment in criminal sentencing. *Annual review of clinical psychology*, *12*, 489-513.

New York State Criminal Procedure Law § 510.30.

New York City Administrative Code §14-151

Oleson, J. C. (2011). Risk in sentencing: Constitutionally suspect variables and evidence-based sentencing. *SMUL Rev., 64*, 1329.

Perkins, K. L. & Sampson, R. J. (2015). Compounded deprivation in the transition to adulthood: The intersection of racial and economic inequality among Chicagoans, 1995–2013. *The Russell Sage Foundation Journal of the Social Sciences, 1*(1), 35-54.

Pettit, B. (2012). *Invisible men: Mass incarceration and the myth of black progress*. New York, NY: Russell Sage Foundation.

Polaschek, D. L. (2012). An appraisal of the risk–need–responsivity (RNR) model of offender rehabilitation and its application in correctional treatment. *Legal and criminological Psychology*, *17*(1), 1-17.

Pretrial Justice Institute (2013, February) Colorado Pretrial Assessment Tool (CPAT): Administration, Scoring, and Reporting Manual, Version 1. *Pretrial Justice Institute.* Retrieved August 2018 from http://capscolorado.org/yahoo_site_admin/assets/docs/CPAT_Manual_v1_-_PJI_2013.279135658.pdf

Richardson, L. S., & Goff, P. A. (2012). Implicit racial bias in public defender triage. *Yale L. J.*, *122*, 2626.

Schlesinger, T. (2013). Racial disparities in pretrial diversion: An analysis of outcomes among men charged with felonies and processed in state courts. *Race and justice*, *3*(3), 210-238.

Schuppe, J. (2017, August 22). Post Bail. *NBC News*. Retrieved from https://www.nbcnews.com/specials/bail-reform

Schwalbe, C. S., Fraser, M. W., Day, S. H., & Cooley, V. (2006). Classifying juvenile offenders according to risk of recidivism: Predictive validity, race/ethnicity, and gender. *Criminal Justice and Behavior*, *33*(3), 305-324.

Simoiu, C., Corbett-Davies, S., & Goel, S. (2016). Testing for racial discrimination in police searches of motor vehicles. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2811449

Soss, J., & Weaver, V. (2017). Learning from Ferguson: Policing, race, and class in American politics. *Annual Review of Political Science*, *20*(1), 565-591.

Spencer, K. B., Charbonneau, A. K., & Glaser, J. (2016). Implicit bias and policing. *Social and Personality Psychology Compass*, *10*(1), 50-63.

Starr, S. B. (2014). Evidence-based sentencing and the scientific rationalization of discrimination. *Stan. L. Rev.*, 66, 803.

Starr, S. B. (2015). The new profiling: Why punishing based on poverty and identity is unconstitutional and wrong." *Federal Sentencing Reporter*, *27*(4), 229–236.

Steffensmeier, D., & Demuth, S. (2000). Ethnicity and sentencing outcomes in U.S. federal courts: Who is punished more harshly? *American Sociological Review*, *65*(5), 705–729.

Steffensmeier, D., Ulmer, J., & Kramer, J. (1998). The interaction of race, gender, and age in criminal sentencing: The punishment cost of being young, black, and male. *Criminology, 36*(4), 763–798.

Stevenson, M.T. (2018). Assessing Risk Assessment in Action. *Minnesota Law Review*, *103*, Forthcoming. Available at SSRN: https://ssrn.com/abstract=3016088 or http://dx.doi.org/10.2139/ssrn.3016088

Stolzenberg, L., D'Alessio, S. J., & Eitle, D. (2013). Race and cumulative discrimination in the prosecution of criminal defendants. *Race and Justice*, *3*(4), 275-299.

Summers, C., & Willis, T. (2010). Pretrial Risk Assessment.

van Eijk, G. (2017). Socioeconomic marginality in sentencing: The built-in bias in risk assessment tools and the reproduction of social inequality. *Punishment & Society*, *19*(4), 463-481.

VanNostrand, M., & Keebler, G. (2009). Pretrial risk assessment in the federal court. *Federal Probation*, 73, 3.

Weaver, R. K. (1986). The politics of blame avoidance. *Journal of Public Policy*, *6*(4), 371–398.

Western, B. (2007). *Punishment and inequality in America*. New York, NY: Russell Sage Foundation.

Ward, G. (2015). The slow violence of state organized race crime. *Theoretical Criminology*, *19*(3), 299-314.

Westervelt, E. (2017, August 18). Did a bail reform algorithm contribute to this San Francisco man's murder? *NPR: All Things Considered*.

Wexler, R. (2017). Life, liberty, and trade secrets: Intellectual property in the criminal justice system. *Stan. L. Rev.,* 70, 1343.

White, M. D. (2014). *Police officer body-worn cameras: Assessing the evidence*. Office of Justice Programs, US Department of Justice.

Wisconsin v. Eric Loomis, 881 N.W.2d 749 (Wis. 2016).

Wooldredge, J., Frank, J., Goulette, N., & Travis III, L. (2015). Is the impact of cumulative disadvantage on sentencing greater for Black defendants?. *Criminology & Public Policy*, *14*(2), 187-223.

Yang, M., Wong, S. C., & Coid, J. (2010). The efficacy of violence prediction: a meta-analytic comparison of nine risk assessment tools. *Psychological bulletin*, *136*(5), 740.

Yokum, D., Ravishankar, A., & Coppock, A. (2017, October 20). Evaluating the effects of police body-worn cameras: A randomized controlled trial. *The DC Lab Working Paper.*

Zliobaite, I. (2017). Fairness-aware machine learning: a perspective. *ArXiv:1708.00754 [Cs, Stat]*. Retrieved from http://arxiv.org/abs/1708.00754