The Scales of (Algorithmic) Justice: Tradeoffs and Remedies

Matthew Sun (Stanford University; mattsun@stanford.edu)
Marissa Gerchick (Stanford University; gerchick@stanford.edu)
DOI: 10.1145/3340470.3340478

Introduction

Every day, governmental and federally funded agencies — including criminal courts, welfare agencies, and educational institutions — make decisions about resource allocation using automated decision-making tools (Lecher, 2018; Fishel, Flack, & DeMatteo, 2018). Important factors surrounding the use of these tools are embedded both in their design and in the policies and practices of the various agencies that implement them. As the use of such tools is becoming more common, a number of questions have arisen about whether using these tools is fair, or in some cases, even legal (*K.W. v. Armstrong*, 2015; ACLU, Outten & Golden LLP, and the Communications Workers of America, 2019).

In this paper, we explore the viability of potential legal challenges to the use of algorithmic decision-making tools by the government or federally funded agencies. First, we explore the use of risk assessments at the pre-trial stage in the American criminal justice system through the lens of equal protection law. Next, we explore the various requirements to mount a valid discrimination claim — and the ways in which the use of an algorithm might complicate those requirements — under Title VI of the Civil Rights Act of 1964. Finally, we suggest the adoption of policies and guidelines that may help these governmental and federally funded agencies mitigate the legal (and related social) concerns associated with using algorithms to aid decision-making. These policies draw on recent lawsuits relating to algorithms and policies enacted in the EU by the General Data Protection Regulation (GDPR) (2016).

Algorithms and Equal Protection

One case of algorithmic decision-making in the public domain that has been recently subjected to increased scrutiny in recent years is the use of risk assessments in the criminal justice system. Here, we focus on the use of criminal risk assessment at the pretrial stage. The goal of risk assessment tools (RATs) at the pre-trial stage is typically to estimate a defendant's likelihood of engaging in a particular future action (for example, committing a new crime or failing to appear in court) based on their similarity to defendants who have committed those actions in the past (Summers & Willis, 2010). This similarity is typically determined using factors regarding a defendant's criminal history but may also include information about a defendant's personal and social history such as their age, housing and employment status, and in some cases, their gender (Summers & Willis, 2010; *State v. Loomis*, 2016). Risk assessments are not themselves decision-makers regarding detention; rather, they are tools used by a human decision-maker - typically a judge or magistrate (Desmarais & Lowder, 2019).

In this section, we explore legal challenges pertaining to risk assessments on the basis that their use, under some circumstances, may violate constitutional protections. In particular, the Fifth Amendment guarantees equal protection under due process of law and applies to the federal government (U.S. Const., amend. V.), while the Fourteenth Amendment guarantees equal protection and due process of law and applies to the states (U.S. Const., amend. XIV.). Our analysis focuses on the application of equal protection law to the use of algorithmic risk assessments. Specifically, we discuss policies around the use of gender and proxies for race in risk assessments and how each might interact with equal protection of the law.

When an individual or entity believes that their right to equal protection has been violated by a governmental policy - such as the use of a risk assessment algorithm at the pretrial stage - they may challenge such a policy by, first, proving that the policy does indeed discriminate in a way that is or was harmful to the indi-

vidual (Legal Information Institute, 2018a).The court evaluating the matter would then analyze the policy in question through one of four possible lenses - strict scrutiny, intermediate scrutiny, rational basis scrutiny, or a combination of the prior three, depending on the characteristic (race, national origin, gender, etc.) in question (Legal Information Institute, 2018a).

One such notable challenge, which we reference in the subsequent discussion, was *State of Wisconsin v. Loomis* (2016), in which Eric Loomis challenged the use of the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) risk assessment to inform a judge's decision about how long his prison sentence would be. Loomis challenged the use of COMPAS on the grounds that it violated his constitutional right to due process because the tool itself was proprietary (in particular, Loomis knew the factors used on the assessment but did not know how each of those factors was weighted and translated into a score, and thus could not challenge its scientific validity), and because the tool used gender as a factor in the assessment (*State v. Loomis*, 2016).

Factor 1: Use of gender

Though many risk assessments used at the pretrial stage in the United States do not include gender as a factor in the calculation of risk scores (Latessa, Smith, Lemke, Makarios, & Lowenkamp, 2009; VanNostrand et al., 2009), some pretrial risk assessments do consider gender, like COMPAS did in the case of Eric Loomis (*State v. Loomis*, 2016). Moreover, evidence indicates that risk assessments may not be equally predictive across genders, and may overestimate the recidivism risk of women compared to men (Skeem, Monahan, & Lowenkamp, 2016). Such evidence suggests the counterintuitive idea that including gender in the calculation of risk scores may be more equitable than excluding it. To illustrate the complexities of this point, we consider two hypothetical scenarios regarding risk assessments and gender.

Consider a hypothetical risk assessment X that includes gender in its calculation of risk scores; assume X has been challenged on the basis that its use of gender violates equal protection. Equal protection claims involving gender classifications are subject to intermediate scrutiny, a test established by the Supreme Court in *Craig v. Boren* (1976). To pass intermediate scrutiny, the policy in question must "advance an important government interest" by means that are "substantially related to that interest" (Legal Information Institute, 2018b; *Craig v. Boren*, 1976). The defendant (the jurisdiction that uses X to inform pretrial release decisions) might argue that, because judges rely on the accuracy of risk scores when making decisions about who to release and because these risk scores are meant to inform their decision-making, the use of gender in X advances an important government interest - ensuring public safety through release determinations. The defendant might also argue that, given the evidence on differential predictive power by gender, the use of gender is indeed a means that is "substantially related" to public safety.

In the case of Loomis, the court determined the use of gender was permissible because it improved accuracy, a non-discriminatory purpose (*State v. Loomis*, 2016). Yet some argue that such evidence regarding the differential predictive power by gender is too general. Legal scholar Sonja Starr has argued that because the Supreme Court has rejected the use of broad statistical generalizations about groups to justify discriminatory classifications, the use of gender in risk assessment (specifically at sentencing) is unconstitutional (Starr, 2014). In the case of X, the court would have to consider, given the relevant evidence, if it is actually the case that using gender as a factor is substantially related to public safety, weighing the tension between the group classifications in X and the principle of individualized decision-making in the criminal justice system.

Now consider risk assessment Y, a risk assessment that doesn't include gender in its calculation of risk scores, and suppose that a jurisdiction that uses Y has analyzed its own data and found that Y is better at predicting recidivism for men than it is at predicting recidivism for women. In this case, the policy in question is facially neutral (the use of Y doesn't appear to be discriminatory towards women and doesn't specifically include gender in its calculations), but nonetheless has a disparate impact because it rates women as higher risk than they actually are. If the use

of Y were challenged under equal protection, the challenger would have to show intent - in particular, that the governmental body using Y intended to discriminate against women by using Y. In *Personnel Administrator of Massachusetts v. Feeney* (1979), the Supreme Court was faced with the question of whether a facially neutral policy that had a disparate impact on women was a violation of equal protection. A key question was whether the "foreseeability" of the policy's disparate impact was sufficient proof of discriminatory intent; the court held that it was not (Weinzweig, 1983). Thus, if the ruling from Feeney were applied to the hypothetical case regarding Y, awareness of Y's differential predictive power for men and women may not necessarily qualify as proof of intent to discriminate, and the equal protection claim against Y may fall short.

Factor 2: Use of proxies for race

Now consider a hypothetical risk assessment Z that uses factors such as the stability of a defendant's housing or their employment status - in practice, many risk assessments do consider these factors, as they are correlated with recidivism risk (Summers & Willis, 2010). However, these factors may serve as proxies for race (Barocas & Selbst, 2016; Corbett-Davies & Goel, 2018). Though classifications involving race or national origin are typically subject to strict scrutiny, absent an explicit discriminatory classification, both disparate impact and discriminatory intent are required to even trigger a scrutiny test (as they would be in the hypothetical case of Y, described above) (*Arlington Heights v. Metropolitan Housing Dev. Corp.*, 1977). Thus, for Z's use to be challenged because of its use of proxies for race, one would need to show both that Z has a disparate impact (for example, that though scores inform decision-making for all people, Z is less accurate for minorities than for white people, which may or may not be true in the case of this hypothetical) and that Z was designed or used to be discriminatory against the minority group(s) in question. Demonstrating this intent may prove challenging because of the correlation between these socioeconomic factors and recidivism risk; nonetheless, the tension between statistical generalizations about groups of people and the right to an individualized decision for each defendant is ever present.

More broadly, legal challenges to the use of RATs under constitutional law speak to an underlying theme of the use of algorithms more generally: the use of these tools does not fit neatly into established legal standards (Barocas & Selbst, 2016), and tradeoffs will be present, whether mathematical, social, both, or otherwise (Corbett-Davies & Goel, 2018). Moreover, in the presence of facially neutral RATs, understanding intent is crucial to understanding if the law has been violated. In the remedies section, we propose inquires around RAT implementation that may help clarify the intent of policymakers and agencies who adopt these tools and inform the public about the agencies' decision-making rationale in the presence of tradeoffs.

Algorithms and Civil Rights Law

Beyond the constitutional arena, disparate impact theory has another, distinct form in civil rights law. Famously, Title VII of the Civil Rights Act of 1964 explicitly bars employment practices that would generate a disparate impact, defined by the following conditions: 1) the policy creates an adverse effect that falls disproportionately upon a particular protected class, 2) the specific policy in place is not a "business necessity," and 3) there exists an alternative policy that would not result in disproportionate harms (42 U.S.C. §2000e et seq.). In *Griggs v. Duke Power Co.* (1971), the Supreme Court found that Duke Power's requirement of a high school diploma for its higher paid jobs was illegal under Title VII of the Civil Rights Act of 1964 because it disproportionately barred minority groups from those positions and did not have any demonstrable relation to performance on the job.

Beyond Title VII and employment practices, the Court has ruled in multiple cases involving federal statutes with disparate impact provisions, such as *Lau v. Nichols* (1974) and *Alexander v. Choate* (1985), that policies which create adverse disparate impact are in violation of the law, regardless of the intent of those policies or whether the policies are applied equally to all groups. Such policies that create a disparate impact constitute a violation of Title VI of the Civil Rights Act of 1964, which was enacted at the same time as Title VII (42 U. S. C. § 2000d). We choose to now

shift focus to Title VI because Title VI stipulates that all programs or activities that receive federal funding may not perpetrate or perpetuate discrimination on the grounds of race, color, or national origin, while Title VII only concerns employment (42 U. S. C. § 2000d). However, we note that the U.S. Department of Justice has recently stated that Title VI "follows...generally..the Title VII standard of proof for disparate impact"; thus, cases that concern Title VII "may shed light on the Title VI analysis in a given situation" (U.S. Department of Justice, 2019).

Twenty-six federal agencies have Title VI regulations that address the disparate impact standard, including USDA, the Department of Health and Human Services, and the Department of Education (U.S. Department of Justice, 2019). These federal agencies provide funding to a massive array of public programs and the social safety net, including public schools, Medicaid, and Medicare. In *Lau v. Nichols* (1974), for example, the Court found that the San Francisco Unified School District was in violation of Title VI because it received federal funding yet imposed a disparate impact on non English-speaking students, many of whom were not offered supplemental language instruction or placed into special education classes.

This regulatory and legal landscape sets the stage for the application of disparate impact theory under civil rights law as an important possible remedy for discrimination in algorithmic decision-making. As state and local governments increasingly turn towards automated tools to lower costs, ease administrative burdens, and deliver benefits, we are likely to observe cases where algorithms, especially when deployed without comprehensive oversight and auditing processes in place, create unequal outcomes. In her book *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*, Professor Virginia Eubanks examines a statistical tool used by the Allegheny County Office of Youth, Children, and Families that processes data from public programs to predict the likelihood that child abuse is taking place in individual households across the county (Eubanks, 2018). Because the frequency of calls previously made on a family is an input to the algorithm, Eubanks argues that the tool may systematically discriminate against Black families, since Black families are far more likely to be called on by mandatory reporters or anonymous callers (Misra, 2018). The Office of Youth, Children, and Families is overseen by the Allegheny County Department of Human Services, which receives federal funding and as a result may be subject to regulation under Title VI (Allegheny County, 2019).

In these cases and many others, there is often no obvious evidence of discriminatory intent; to the contrary, algorithms are commonly deployed in the hopes of mitigating human biases (Lewis, 2018). In Allegheny County, officials stressed that the predictive risk-modeling tool would guide, not replace, human decision-making (Hurley, 2018; Giammarise, 2017). Yet, we often see that algorithms may still produce significant adverse impact on populations when analyzed on the basis of race or gender. As a result, groups or individuals may naturally seek to challenge the use of such algorithms in programs receiving federal funding under Title VI. According to a Justice Department legal manual on Title VI, three conditions are required to constitute a violation of Title VI: 1) statistical evidence of disparate adverse impact on a race, color, or national origin group, 2) the lack of a substantial legitimate justification for the policy, and 3) the presence of a less discriminatory alternative that would achieve the same objective but with less of a discriminatory effect (42 U. S. C. § 2000d).

In the following sections, we explore how disparate impact claims against the usage of algorithms might fail to succeed in court for three separate reasons. These challenges can be summarized as the lack of presence of a less discriminatory alternative, the use of predictive accuracy as "substantial legitimate justification" for the policy, and the possibility that the only way to ameliorate disparate impact would be to treat different groups differently, thus triggering a disparate treatment legal challenge. We explore the current standard for how a complainant (i.e., plaintiff) must prove disparate impact under Title VI, and how a recipient (i.e., defendant) might ultimately circumvent their claims.

Challenge 1: Proving the presence of a less discriminatory alternative

The phrase "less discriminatory alternative"

implies that there exists a way to compare a set of policies and determine which is the least discriminatory. However, when it comes to algorithmic decision-making, the definition of "fairness" (in other words, the absence of discrimination) is hotly debated (Gajane & Pechenizkiy, 2017). For example, the notion of "classification parity" is defined as the requirement that certain measures of predictive performance, such as the false positive rate, precision, and proportion of decisions that are positive, be equal across protected groups (Corbett-Davies & Goel, 2018). For example, in order to satisfy false positive classification parity, the Allegheny County child neglect prediction algorithm must make an incorrect positive prediction (i.e., predict the presence of child abuse in a family where none is occurring) at the same rate for both White and Black families. Another commonly referenced notion of fairness is "calibration," which requires that outcomes be independent of protected class status after controlling for estimated risk (Corbett-Davies & Goel, 2018). If the aforementioned algorithm were to satisfy calibration, child abuse must be found to actually occur at similar rates in White and Black families predicted to have a 10% risk of child neglect.

These definitions may sound like they measure roughly similar phenomena, but recent research on algorithmic fairness shows that they are often in competition, producing provable mathematical tradeoffs among each other (Corbett-Davies & Goel, 2018). Optimizing calibration, for example, may result in reductions in classification parity. ProPublica's analyis of the use of COMPAS at the pretrial stage in Broward County, Florida revealed that the algorithm yielded much higher false positive rates for Black defendants than it did for White ones (Angwin, Larson, Mattu, & Kirchner, 2016), but at the same time, individuals given the same COMPAS risk score recidivated at the same rate (Corbett-Davies, Pierson, Feller, Goel, & Huq, 2017). In other words, the algorithm was calibrated, but was more likely to incorrectly classify Black defendants as "high risk" for recidivism than White defendants. To further complicate the notion of discrimination, the algorithm used in Allegheny County to predict risk of child neglect was miscalibrated in a way that disfavored White children: White children who received the same risk score for neglect as Black children were actually less likely to be experiencing maltreatment (Chouldechova, Benavides-Prado, Fialko, & Vaithianathan, 2018). In this case, Eubanks' critiques of the algorithm's inputs and other researchers' empirically measured calibration result in directly opposing views of which racial group is experiencing discrimination.

Without a single, legally-codified definition of fairness, we see the first obstacle to a successful disparate impact claim: a recipient can argue that no less discriminatory alternative exists, since any alternative will likely involve tradeoffs across different measures of fairness. Moreover, we suggest that it is insufficient to choose one measure of fairness as the priority in all cases, since the societal costs associated with different fairness measures varies across specific applications (Corbett-Davies & Goel, 2018). For example, one might argue that the societal and/or moral cost of incorrectly detaining a Black individual who will not recidivate is far greater than the cost of incorrectly flagging a Black household for child abuse. Another person might take the opposite position, but in either case, blindly prioritizing false positive parity across both tasks would fail to recognize the unique costs associated with each one.

There also exist practical legal challenges and ambiguity regarding the existence of a less discriminatory alternative. In the realm of Title VII, scholars disagree about whether "refusal" to adopt a less discriminatory procedure means that the employer cannot be held liable until it has actively investigated such an alternative and subsequently rejected it (Barocas & Selbst, 2016). This debate raises the question of whether employers should be held responsible to perform a costly, exhaustive search of all potential alternatives, or whether the cost of doing such a search would functionally mean that less discriminatory alternatives do not exist. According to the U.S. Department of Justice's guidance regarding Title VI, the burden is on the complainant to identify less discriminatory alternatives (U.S. Department of Justice, 2019). This may pose a significant challenge to complainants, as they may not have access to the documents and data needed to show which alternatives would be equally ef-

fective in practice.

Challenge 2: Substantial legitimate justification

The second failure mode for a disparate impact claim is that the recipient has articulated a "substantial legitimate justification" for the challenged policy (42 U. S. C. § 2000d). As the Justice Department discloses in its Title VI legal manual, "the precise nature of the justification inquiry in Title VI cases is somewhat less clear in application" (U.S. Department of Justice, 2019). For example, the EPA stated in its 2000 Draft Guidance for Investigating Title VI Administrative Complaints that the "provision of public health or environmental benefits...to the affected population" was an "acceptable justification" (Draft Title VI Guidance, 2000). This document was compiled after a 60-day period of 7 public listening sessions at the request of state and local officials seeking clarification in an effort to avoid Title VI violations (Mank, 2000). In contrast, Title VII substitutes the "legitimate justification" requirement with a "business necessity" stipulation (42 U. S. C. § 2000d). Because Title VI covers a broad scope of federally funded programs, "legitimate justification" must be defined on a case-by-case basis, whereas "business necessity" has a narrower meaning in case law due to Title VII's specific focus on hiring practices (U.S. Department of Justice, 2019).

In the case of programmatic decision-making, discrimination may occur when practitioners do not properly audit their algorithm before and while it is deployed. Such an audit could take many forms, such as running a randomized control trial before permanently implementing an algorithm or releasing public reports every year regarding how well the algorithm is performing. (For the purposes of the following discussion, we assume that the task at hand is one of binary/multiclass classification, also known as a "screening procedure"). In the field of machine learning, algorithms are commonly trained by iteratively improving performance on a given dataset, as measured by average classification accuracy (Alpaydin, 2009). If average classification accuracy is not disaggregated across protected groups present in the dataset, disparities in the algorithm's performance may only be discovered once the algorithm is already deployed for real-world use (Buolamwini & Gebru, 2018), which could result in a subsequent disparate impact claim. In this sequence of events, the potentially offending entity was optimizing for overall accuracy and failed to take the possibility of disparate impact into account.

This scenario raises the question of whether the desire to optimize raw predictive accuracy counts as a "substantial legitimate justification" for an algorithm whose outputs are biased. It seems plausible that any recipient could argue that predictive accuracy is a legitimate justification: after all, optimizing accuracy maximizes the total number of decisions made correctly, given that the demographic makeup of the dataset resembles that of the real-world population. Optimizing for any other metric, such as an arbitrary fairness measure, may lead to an algorithm with lower overall predictive accuracy (Zliobaite, 2015; Kleinberg, Mullainathan, & Raghavan, 2016). A recipient of a disparate impact claim could argue that maximizing accuracy leads to higher efficiency and lower costs for cash-strapped government agencies. In the Allegheny County example, having an algorithm accurately flag families for risk of child neglect reduced the time required to manually screen applications, saving time and labor. Because "substantial legitimate justification" is relatively ambiguous and case-specific, it may be difficult for a complainant to prove that maximizing classification accuracy is not a legitimate justification.

Challenge 3: A disparate impact and disparate treatment Catch-22

It's important to note that optimizing accuracy and fairness measures is not always a zero-sum game. In the aforementioned research about gender in criminal risk assessment, including gender as a variable in the dataset improved calibration and predictive accuracy because women with similar criminal histories to men recidivate at lower rates (Skeem et al., 2016) (Notably, gender is not a protected attribute under disparate impact clauses in civil rights law). Similarly, in other cases, we may be able to improve predictive accuracy and produce gains in fairness measure(s) if some predictive latent variable is identified and included in the dataset (Jung, Corbett-Davies, Shroff, & Goel, 2018).

Consider the case of a hypothetical algorithm that estimates recidivism risk and takes race as an input, but does not take criminal history as an input. Assume in this scenario that criminal history is more predictive of recidivism than race. If Black people are disproportionately likely to have prior convictions - perhaps due to disparate policing practices - then the algorithm will "penalize" all Black people by giving them higher risk scores, even ones without prior convictions. If criminal history is added to the dataset and the algorithm is retrained, the algorithm's accuracy will increase due to the addition of a predictive variable. In addition, the algorithm's performance on fairness measures may increase as well, since Black people without criminal histories will no longer receive a penalty for their racial status.

It may be the case, however, that the latent variable whose inclusion would improve fairness and accuracy is the protected attribute itself (Jung et al., 2018). Including gender as an input to the algorithm would resolve the unequal outcomes in which women are unfairly penalized, but at the same time, explicitly altering decisions based off of an individual's gender is a clear example of disparate treatment (42 U. S. C. § 2000d). The same would be true with regard to protected attributes under Title VI such as race, national origin, and religion. Disparate treatment, in which policies explicitly treat members of different protected groups differently, is prohibited by Title VI, as well as many other civil rights laws (U.S. Department of Justice, 2019). Disparate treatment cases are arguably easier to prove, since discrimination is explicitly codified in a recipient's policies, while disparate impact cases rely on measures of a policy's outcomes de facto (Selmi, 2005). The fact that both disparate treatment and disparate impact violate civil rights statutes may create a Catch-22 for entities seeking to resolve disparate impact in algorithmic decision-making.

Indeed, Kroll et al. (2016) note this tension as manifested in the Supreme Court's decision in a 2009 case involving Title VII, *Ricci v. DeStefano* (2009). In the case, the New Haven Civil Service Board (CSB) refused to certify the results of a facially neutral test for firefighter promotions out of disparate impact concerns, noting that the pass rate for minorities was half that for whites. As Kroll et al. (2016) note, the Court's decision to rule *against* the CSB "demonstrates the tension between disparate treatment and disparate impact," since a neutral policy can create disparate outcomes, but mitigating the disparate impact would require discriminatory treatment of different groups.

Remedies

As we have seen from the above analysis, there is reason to believe that today's concerns regarding algorithmic bias will not be resolved in the courts alone, despite the high number of pending court cases regarding the use of algorithms. In the Constitutional realm, absent a suspect classification, both disparate impact and discriminatory intent are needed to prove a violation of the law. In addition, the current requirements to make a successful claim of disparate impact under civil rights law are vague with regards to defining what a discriminatory outcome is, which may allow recipients of complaints to leverage whichever mathematical constructs of fairness best support the use of their algorithm.

If we cannot expect to find remedies from the judiciary, where should citizens turn for relief? To address the above concerns, we propose a remedy in the form of a unified, collaborative effort between the agencies and legislatures, both at the federal and state levels. We detail what such an effort would look like below, using an international regulation to inform our proposals.

The European Union's General Data Protection Regulation (GDPR) offers a compelling case for broad legal regulations coupled with significant enforcement power. The GDPR provides strong protections for individual privacy by allowing governmental agencies to pursue fines and investigations into private companies for data mismanagement and privacy breaches (Steinhardt, 2018). With regards to automated decision-making, the GDPR (2016) makes mention of a "right to explanation" for users who seek explanation for decisions made about them (e.g., loan denials) (Goodman & Flaxman, 2017). One of the European Commission's senior advisory bodies on data protection released a set of guidelines regarding automated decision-making, which included requirements for companies to provide explanations for how users'

personal data was used by the algorithm (Casey, Farhangi, & Vogl, 2018). The same body even included a recommendation for companies to introduce "procedures and measures to prevent...discrimination" and to perform "frequent assessments...to check for any bias" (17/EN. WP 251, 2017).

The fact that the mandates behind the GDPR have been enforced in practice leads us to suggest an approach in the U.S. that similarly combines comprehensive legislation with new enforcement powers for government agencies (Lawson, 2019). Of course, attitudes and policies regarding the regulation of private companies differ in the U.S. and the EU (Hawkins, 2019). Thus, our proposal would not seek to impose regulations on all private companies across the US, but rather public entities that are already subject to significant government oversight, such as federal agencies or federally-funded programs. Indirectly, this implicates private companies such agencies may contract with to provide tools or services in their use of algorithmic technology.

The remedies we suggest apply to both of the main use cases we previously described; for federally funded agencies, these remedies may be enacted through legislation or executive rule-making. Similarly, these remedies could also be applied at the state and local level. In both cases, we recommend the creation or significant expansion of agencies focused specifically on the technical oversight and evaluation of algorithmic tools. For example, such an existing agency that might take up this burden could be the newly created Science, Technology Assessment, and Analytics team at the U.S. Government Accountability Office (U.S. Government Accountability Office, 2019). While courts have been reluctant to conduct a "searching analysis of alternatives," federal agencies are "subject matter experts charged with Title VI enforcement duties" and "are well-equipped to...evaluate carefully potential less discriminatory alternatives" (U.S. Department of Justice, 2019).

Our remedy additionally attempts to recognize and address the significant gap in current civil rights legislation with regards to definitions of discriminatory intent and disparate impact — which can generate a Catch-22 of sorts, even for well-meaning actors. Existing civil rights legislation largely focuses on barring discriminatory intent that results in differential treatment on the basis of protected attributes, such as race. Today, however, we see that in order to remedy unintended discrimination in algorithmic decision-making, we may have to take into account such protected attributes: essentially, using differential treatment to ameliorate disparate outcomes. Federal and state legislation must acknowledge this nuance, allowing practitioners to use protected attributes data to promote the most fair outcomes, where the relevance of such data and a suitable notion of fairness are determined on a case by case basis. For example, under a bail reform law in New Jersey, agencies may collect information about a defendant's race and gender for potential use in a risk assessment calculation, subject to the condition that decisions are not discriminatory along race or gender lines (NJ Rev Stat § 2A:162-25, 2014).

Legislation (or other regulation) should stipulate that public agencies that are going to adopt algorithms to help make decisions must submit the following information to a relevant oversight agency (at the federal level, the office described earlier, and at the state level, some state or local agency with relevant expertise) prior to the algorithm's adoption:

- *What decision will the algorithm be used to make or help make?* How was that decision or type of decision made before the use of the algorithm?
- *What are the reasons to implement such an algorithm?* Is the algorithm less expensive, or will it increase efficiency? Is the intent to make the decision-making process more objective?
- *What are the particular use cases and use context of the algorithm?* How will the algorithm's outputs be interpreted? Will a human decision-maker be involved? Who is the population that the algorithm may be used on? Are there any exceptions to this policy?
- *How will the algorithm be evaluated and, if necessary, revised?* Has funding been allocated for regular oversight? Who will be performing the evaluations and how? Is the text (or source code) or training data of the algorithm publicly available?
- *Were alternatives considered?* What other options were considered, and why was this

one chosen? What were tradeoffs between the different choices?

The submission of this information to a governmental body and the public before an algorithm is employed in practice could provide greater clarity to both the public and regulators regarding discriminatory intent and the potential for discriminatory outcomes. Furthermore, by actively requiring actors to come up with a plan to monitor the algorithm, consider alternatives, and think critically about the algorithm in the context of human systems, this policy may decrease the likelihood of algorithms producing unintended negative consequences in practice.

Acknowledgments

We would like to thank Keith Schwarz for helpful feedback.

References

17/EN. WP 251. (2017). *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679*

ACLU, Outten & Golden LLP, and the Communications Workers of America. (2019). *Facebook EEOC complaints.* https://www.aclu.org/cases/facebook-eeoc-complaints. (Online; accessed June 1, 2019)

Alexander v. Choate, 469 U.S. 287 (1985)

Allegheny County. (2019). *DHS funding.* https://www.county.allegheny.pa.us/Human-Services/About/Funding-Sources.aspx. (Online; accessed May 18, 2019)

Alpaydin, E. (2009). *Introduction to machine learning*. MIT press.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. *ProPublica, May*, *23*.

Arlington Heights v. Metropolitan Housing Dev. Corp., 429 U.S. 252 (1977)

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *Calif. L. Rev.*, *104*, 671.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77–91).

Casey, B., Farhangi, A., & Vogl, R. (2018). Rethinking explainable machines: The GDPR's "right to explanation" debate and the rise of algorithmic audits in enterprise.

Chouldechova, A., Benavides-Prado, D., Fialko, O., & Vaithianathan, R. (2018). A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency* (pp. 134–148).

Civil Rights Act of 1964 Title VI, 78 Stat. 252, 42 U. S. C. § 2000d.

Civil Rights Act of 1964 Title VII, 42 U.S.C. §2000e et seq.

Corbett-Davies, S., & Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining* (pp. 797–806).

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

Desmarais, S. L., & Lowder, E. M. (2019). Pretrial risk assessment tools: A primer for judges, prosecutors, and defense attorneys. *MacArthur Foundation Safety and Justice Challenge*.

Draft Title VI Guidance for EPA Assistance Recipients Administering Environmental Permitting Programs (Draft Recipient Guidance) and Draft Revised Guidance for Investigating Title VI Administrative Complaints Challenging Permits (Draft Revised Investigation Guidance); Notice, 65 Fed. Reg. 124 (June 27, 2000). Federal Register: The Daily Journal of the United States.

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

Fishel, S., Flack, D., & DeMatteo, D. (2018). Computer risk algorithms and judicial decision-making. *Monitor on Psychology*.

Gajane, P., & Pechenizkiy, M. (2017). On formalizing fairness in prediction

with machine learning. *arXiv preprint arXiv:1710.03184*.

Giammarise, K. (2017). *Allegheny County DHS using algorithm to assist in child welfare screening.* https://www.post-gazette.com/local/region/2017/04/09/Allegheny-County-using-algorithm-to-assist-in-child-welfare-screening/stories/201701290002. (Online; accessed May 18, 2019)

Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, *38*(3), 50–57.

Griggs v. Duke Power Co., 401 U.S. 424 (1971)

Hawkins, D. (2019). *The cybersecurity 202: Why a privacy law like gdpr would be a tough sell in the U.S.* https://www.washingtonpost.com/news/powerpost/paloma/the-cybersecurity-202/2018/05/25/the-cybersecurity-202-why-a-privacy-law-like-gdpr-would-be-a-tough-sell-in-the-u-s/5b07038b1b326b492dd07e83/?utm_term=.1cc41e57f9cf. (Online; accessed May 18, 2019)

Hurley, D. (2018). Can an algorithm tell when kids are in danger. *New York Times*, *2*.

Jung, J., Corbett-Davies, S., Shroff, R., & Goel, S. (2018). Omitted and included variable bias in tests for disparate impact. *arXiv preprint arXiv:1809.05651*.

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.

K.W. v. Armstrong, No. 14-35296 (9th Cir. 2015)

Latessa, E., Smith, P., Lemke, R., Makarios, M., & Lowenkamp, C. (2009). Creation and validation of the Ohio risk assessment system: Final report. *Cincinati, OH: University of Cincinnati*.

Lau v. Nichols, 414 U.S. 563 (1974)

Lawson, R. P. (2019). *GDPR enforcement actions, fines pile up.* https://www.manatt.com/Insights/Newsletters/Advertising-Law/GDPR-Enforcement-Actions-Fines-Pile-Up. (Online; accessed May 18, 2019)

Lecher, C. (2018). What happens when an algorithm cuts your health care. *The Verge*.

Legal Information Institute. (2018a). *Equal protection.* https://www.law.cornell.edu/wex/equal_protection. (Online; accessed May 18, 2019)

Legal Information Institute. (2018b). *Intermediate scrutiny.* https://www.law.cornell.edu/wex/intermediate_scrutiny. (Online; accessed June 1, 2019)

Lewis, N. (2018). *Will AI remove hiring bias?* https://www.shrm.org/resourcesandtools/hr-topics/talent-acquisition/pages/will-ai-remove-hiring-bias-hr-technology.aspx. (Online; accessed May 18, 2019)

Mank, B. C. (2000). The draft recipient guidance and the draft revised investigation guidance: Too much discretion for epa and a more difficult standard for complainants? *Environmental Law Reporter*, *30*.

Misra, T. (2018). *When criminalizing the poor goes high-tech.* https://www.citylab.com/equity/2018/02/the-rise-of-digital-poorhouses/552161/?platform=hootsuite. (Online; accessed May 18, 2019)

NJ Rev Stat § 2A:162-25 (2014)

O.J. (L 119) (2016). *Reg (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Dir 95/46/EC (General Data Protection Regulation)*

Personnel Adm'r of Massachusetts v. Feeney, 442 U.S. 256 (1979)

Reg (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free move-

ment of such data, and repealing Dir 95/46/EC (General Data Protection Regulation). (2016).

Ricci v. DeStefano, 557 U.S. 557 (2009)

Selmi, M. (2005). Was the disparate impact theory a mistake. *Ucla L. Rev.*, *53*, 701.

Skeem, J., Monahan, J., & Lowenkamp, C. (2016). Gender, risk assessment, and sanctioning: The cost of treating women like men. *Law and human behavior*, *40*(5), 580.

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

State v. Loomis, 881 N.W.2d 749 (2016)

Steinhardt, E. (2018). *European regulators are intensifying GDPR enforcement.* https://www.insideprivacy.com/eu-data-protection/european-regulators-are-intensifying-gdpr-enforcement/. (Online; accessed May 18, 2019)

Summers, C., & Willis, T. (2010). Pretrial risk assessment research summary. *Washington, DC: Bureau of Justice Assistance*.

United States. Department of Justice. (2019). *Title VI Legal Manual (Updated)*

U.S. Government Accountability Office. (2019). *Our new science, technology assessment, and analytics team.* https://blog.gao.gov/2019/01/29/our-new-science-technology-assessment-and-analytics-team/. (Online; accessed May 18, 2019)

U.S. Const. amend. V.

U.S. Const. amend. XIV.

VanNostrand, M., et al. (2009). Pretrial risk assessment in Virginia.

Weinzweig, M. J. (1983). Discriminatory impact and intent under the equal protection clause: The Supreme Court and the mind-body problem. *Law & Ineq.*, *1*, 277.

Zliobaite, I. (2015). On the relation between accuracy and fairness in binary classification. *arXiv preprint arXiv:1505.05723*.

Marissa Gerchick is a rising senior at Stanford University studying Mathematical and Computational Science. She is interested in using data-driven tools to improve the American criminal justice system.

Matthew Sun is a rising senior at Stanford double majoring in Computer Science and Public Policy. He leads a student group called CS+Social Good and is interested in applied AI research for socially relevant issues.