Should Algorithms that Predict Recidivism Have Access to Race?

1. Introduction

In 2016, ProPublica published a bombshell report on what had been a relatively obscure yet increasingly common practice within the American criminal justice system: the use of algorithmic risk assessments to inform decision making about sentencing, bail, and parole (Angwin et al. 2016). The report focused specifically on the disparity in risk scoring errors between white and black defendants produced by the Correctional Offender Management Profile for Alternative Sanctions (COMPAS) algorithm. COMPAS, which is used widely across the United States, assigns defendants risk scores ranging from 1 to 10. Each score reflects the system's assessment of the probability that the defendant will reoffend within two years of release. In a pair of cases featured by ProPublica, a Black woman and a white man were each arrested for (the very same offense of) petty theft. The woman had previously been charged with four misdemeanors as a juvenile; the white man had previously been charged with two armed robberies and one attempted armed robbery. And yet, the algorithm classified the white man as 'low risk' and the Black woman as 'high risk'. These classifications were mistaken; the Black woman ultimately did not go on to reoffend, while the white man did: he is currently serving eight years for grand theft.

This was not a one-off result. The report showed both that Black defendants are roughly twice as likely as white defendants to be mistakenly classified as medium- or high-risk, and that white defendants who re-offended were erroneously classified as low-risk almost twice as often as Black defendants who re-offended. Since the publication of the ProPublica report, other similar cases have come to light, in which Black defendants are more likely to be mistakenly classified by recidivism risk assessment algorithms as posing a higher risk of reoffending than their white counterparts. For example, in December 2021 the Department of Justice revealed that a criminal risk assessment

system used to make decisions about eligibility for early release programs demonstrated error rate disparities between white and non-white federal inmates (NIJ 2021). There are several initially plausible explanations for these disparities—differences in past criminal activity, age, prior recidivism, etc.; however, the disparities persist even when these factors are controlled for. The natural conclusion to draw, then, is that these algorithms are in some way biased in a way that is to the disadvantage of Black defendants in comparison with white defendants.

This disparity in error rates between white and Black defendants strikes many as unfair. ProPublica's evidence that COMPAS is unfair is that it violates a statistical requirement of algorithmic fairness known as *error parity*. Roughly put, error parity is a standard of fairness according to which an algorithmic system is fair only if its false positive rate and false negative rate are equal across protected groups. A false positive occurs when an algorithmic system incorrectly classifies a subject as possessing some feature of interest. A false negative occurs when an algorithmic system incorrectly classifies a subject as not possessing that feature. An algorithm that violates error parity, say, by producing many more false positives for one group than another, displays a bias concerning one group in relation to the other. This is especially ethically worrisome in cases where algorithmic classification can lead to harmful consequences for a person such as pretrial detention. COMPAS clearly violated error parity because the false positive rate for Black defendants was higher than the false positive rate for white defendants—in the context of risk assessment scores this means that Blacks were mistakenly deemed high risk more often than whites. There remains disagreement about whether violations of error parity constitute unfairness and whether COMPAS in fact violated error parity in the first place.[1] Nevertheless, our interest in this paper is to determine what it is permissible

[1] For arguments for the position that a violation of error parity (but not calibration) is evidence of unfairness see: (Castro 2019; Hellman 2020). For discussion of this case, and the challenge of disagreement about fairness metrics, see: (Corbett-Davies et al. 2016). For a recent argument against

to do, if we assume that the (apparent) violation of error parity by COMPAS is seriously morally objectionable because it is unfair.[2]

In response to the increased attention to apparent violations of fairness by criminal risk assessment algorithms, many activists and scholars have called for an end to their use in the criminal justice system (Heaven 2020). But another possibility is that risk assessment algorithms could be improved along the dimension of fairness while preserving their utility as a risk prediction tool. Recently, some scholars have suggested that error parity might be achieved by giving the algorithm access to a defendant's race, which would allow it to identify other traits that are more predictive of recidivism for individuals of that particular group (Skeem and Lowenkamp 2020; Hellman 2020; Huq 2019; Corbett-Davies et al. 2017). For example, if housing stability were less predictive of recidivism in Blacks than whites, then housing stability would be included in the algorithmic assessment for white defendants, but not for Black defendants. This could potentially reduce the burden for Black defendants while improving accuracy overall. In a recent paper, Jennifer Skeem and Christopher Lowenkamp found that allowing algorithms access to race in this way effectively preserved the predictive value of the algorithms while minimizing imbalances in error rates across racial groups (Skeem and Lowenkamp 2020, 259).[3]

---

the use of error parity to assess algorithmic fairness, see: (Hedden 2021). Whether one finds COMPAS unfair therefore ultimately depends on which approach to algorithmic fairness one favors.

[2] In response to ProPublica's argument, several critical responses have been issued. First, the developers of COMPAS have argued that ProPublica's analysis involved several important statistical and technical errors, which when corrected, demonstrate COMPAS's lack of bias (Equivant 2018). Others have criticized ProPublica's focus on error parity, noting their tendentious approach categorizing errors—namely, by collapsing three categories into two groupings, which generates inaccurate classifications of errors (Flores, Bechtel, and Lowenkamp 2016). See also (Corbett-Davies et al. 2016).

[3] It is unclear whether Skeem and Lowenkamp considered merely a baseline algorithm for white defendants, or one that was specifically tailored to elements with predictive value for white defendants. Our discussion later focuses on the latter possibility, but we are cautious not to attribute that particular approach to Skeem and Lowenkamp here.

The approach just suggested effectively creates different 'tracks' within an algorithm for different racial groups. Each track takes different features into account when producing a risk score–or applies different weights to those features–depending on the feature's predictive value for that racial group. For the sake of the discussion to follow we assume that 'creating different tracks for different racial groups' means departing from a default algorithmic system which takes the same recidivism-relevant features into account regardless of a defendant's race, or that applies equal weight to those features regardless of race. If the two groups we are evaluating are white and Black defendants, then implementing different racial tracks will involve moving *both* white and Black defendants from the default system onto different tracks each of which includes only features with the greatest predictive value for members of that group.

A different race-sensitive approach to risk assessment uses a defendant's race to determine the risk *threshold* or *cut point* to apply to them. For example, a race-neutral algorithm might label anyone 'high risk' who receives a score of six or higher. On the other hand, an algorithm that is sensitive to the defendant's race might label any white defendant 'high risk' who receives a score of six or higher, while labeling a Black defendant 'high risk' if and only if they receive a risk score of eight or higher.

This essay analyzes these two methods of including race in algorithmic sentencing—i.e., different racial tracks and different racial cut points—to better understand whether there is any moral difference between them. Some scholars have assumed that including race in either of these ways risks violating equal treatment under the law (Corbett-Davies et al. 2017). In contrast, Deborah Hellman has recently argued that while implementing different cut points would likely constitute disparate treatment and thus be legally prohibited, employing different tracks for white and Black defendants could be legally justified (Hellman 2020, 852–55). In section 2 we argue that Deborah Hellman fails to establish a moral distinction between using different racial cut points and using

different racial tracks and that she also probably fails to establish a legal distinction between them. We conclude that the conditions for disparate treatment identified by Hellman support the conclusion that if the use of different racial cutpoints constitutes disparate treatment, then so does the use of different racial tracks. In section 3 we identify several previously underappreciated moral differences between these two methods of achieving error parity. We argue that these differences may vindicate the use of different racial tracks—by enabling the practice to withstand strict scrutiny—even if the practice constitutes disparate treatment.

Before turning to the details of Hellman's argument, two points of clarification are in order. First, whereas Hellman draws a conclusion about the legal permissibility of giving algorithmic systems access to race, we are primarily interested here in the moral permissibility of giving algorithmic systems access to race. Hellman argues that the use of different racial tracks does not constitute disparate treatment. She further argues that because it does not constitute disparate treatment it does not threaten the legal right to equal protection of the law. Therefore, the use of different racial tracks does not trigger strict scrutiny. This is a conclusion about what the law requires. But the legal right to equal protection of the laws is important presumably because it protects an important moral right to equal protection of the law. So, Hellman's conclusion has moral as well as legal significance. For if the use of different racial tracks violates the legal right to equal protection, this poses a prima facie threat to the moral right to equal protection. Even if we suppose that legal rights violations are not coextensive with corresponding moral rights violations, a violation of the former right is at least compelling evidence of a violation of the latter right. There is thus an important conceptual connection between the legal arguments presented by Hellman and the moral issues that are the focus of this paper. Second, following Hellman and others, our focus in this essay is on comparisons between Black and white defendants. This is not to suggest that related issues do

not arise for other racial groups, such as Latinx/Hispanic defendants, or between male and female defendants.

2. Hellman's Argument

Hellman's argument centers on the thought that certain ways of appealing to race in judicial decision-making are likely to be challenged as a form of "disparate treatment" and thus run afoul of the Equal Protection Clause of the Fourteenth Amendment.[4] This, in turn, would likely trigger strict scrutiny—an elevated legal standard which requires that the state demonstrate a "compelling governmental interest," and which is difficult to satisfy in many cases (Hellman 2020, 851).

Hellman takes it for granted that using different cut points for different racial groups constitutes disparate treatment and is thus legally prohibited. Her interest is in whether any other use of race in an algorithmic system might be permitted for the sake of achieving error parity. Hellman argues that it is a mistake to assume that the use of different racial tracks—or, perhaps, related statistical models that depend on the defendant's race—constitutes disparate treatment. Here argument depends on the claim that using different racial tracks fails to satisfy two conditions that the courts have suggested must be met in order for a racial classification to constitute disparate treatment: the effects of the racial classification on affected individuals must be causally direct (rather than remote), and the racial classification must rely on a racial generalization. Because using different racial tracks within an algorithm does not meet these requirements, she argues, it does not constitute disparate treatment. If this argument is sound, then designers of predictive algorithms

[4] This is a good opportunity to stress that, as with virtually any legal argument, it depends heavily on the jurisdiction. In this case, Hellman's legal argument is rooted in the American legal context; use of these algorithmic systems in other countries might not raise these particular legal issues.

have a greater array of methods at their disposal to achieve accuracy and error rate parity in AI systems.

2.1 Racial classifications and remote v. direct effects

Many governmental agencies use racial classifications when collecting information about racial disparities; the US Census is the most comprehensive practice by which the state collects information partly on the basis of racial classifications. The US Census deploys racial categories as part of its information gathering about the US population, and this use has been deemed unproblematic by the courts because, as Hellman puts it, "collection of information is different from use" (Hellman 2020, 857).[5] According to Hellman, the reason that collecting information about race is different from use is that the "real world effects" of information collection about race are causally remote from the racial classifications used to gather Census information:

> [T]he census example suggests that the effect of the racial classification must be direct and not merely the downstream consequence of such classification. The collection of racial data on the Census is highly consequential, after all, with substantial impact in the real world, including for redistricting and for the allocation of governmental resources. And yet, these effects are insufficient to make racial classifications in the Census subject to strict scrutiny. The reason, one suspects, is that these effects are too remote. (Hellman 2020, 858)

So, in order for racial classification to count as disparate treatment, the causal effects of that classification must be in some sense *direct* rather than remote. What counts as a direct rather than

[5] Here Hellman cites the relevant court case (Morales v. Daley 2000).

remote effect is left unclear, but the basic idea appears to be that disparate treatment occurs only if there are very few intervening causal factors between the racial classification and the effect on an individual.

2.2 Racial classifications that do not rely on racial generalizations

Even when the effects of racial classification on individuals are direct rather than remote, this can be legally permitted, Hellman argues, when the racial classification does not itself rely on a racial generalization. What exactly is the distinction between racial classifications that make use of a racial generalization and those that do not? Hellman illustrates the distinction by contrasting the use of suspect descriptions with the practice of racial profiling.

The use of suspect descriptions by police in the course of a criminal investigation does not constitute disparate treatment, because deciding whom to investigate on the basis of a suspect description that includes a description of the suspect's race does not rely on any racial *generalization*, even though it relies on a racial *classification*.[6]

Racial profiling involves using race as at least one factor in determining whether or not to investigate someone or in determining how thoroughly to investigate someone. Unlike the use of suspect descriptions, racial profiling relies on a *statistical generalization* about the probability of a given member of some racial group having committed some crime in comparison with a given member of a different racial group. By relying on a statistical generalization about members of a racial group, racial profiling relies on a racial generalization (Hellman 2020, 859–60). The use of suspect descriptions that include a description of a suspect's race does not rely on any statistical

[6] To be sure, police do rely on a generalization of some sort—one about the reliability of suspect descriptions that include a description of a suspect's race. But they do not rely on a generalization about the suspect's race.

generalizations about members of a racial group. It therefore does not constitute disparate treatment, even though it makes use of racial classifications.

From the example of the U.S. Census and the use of suspect descriptions Hellman derives two individually necessary and jointly sufficient conditions under which the use of racial classifications constitutes disparate treatment. First, the use of racial classifications must have a direct, rather than a remote, effect on individuals. Second, the use of racial classifications must rely on a racial generalization.

2.3 Different racial tracks within algorithms: No proximate effect?

Upon applying this pair of conditions for disparate treatment to the use of different tracks within an algorithm, Hellman infers that the use of different tracks within algorithms does not constitute disparate treatment. As she puts it:

> First, the effect produced by this use of a racial classification is not proximate. Rather, the use of race determines what other factors to employ in making a prediction about recidivism risk. The racial category provides information that in turn can be used to determine what other traits should be incorporated into the algorithm. Like the racial information in the Census, this racial information is likely to have downstream consequences, but these effects are too remote from the use of the classification itself to constitute disparate impact on the basis of race. (Hellman 2020, 862)

Here Hellman suggests that the use of different racial tracks within an algorithm involves making a racial classification that is too far "upstream," causally speaking, for that classification to have a direct effect on any person. Thus it does not constitute disparate impact.

But whether the relevant effects of the racial classification are direct or remote seems to depend on when in the process the classification occurs. In particular, it depends on whether we are considering the *development* of the algorithm or the *deployment* of it in a specific instance. At the developmental phase, the decision about which features to include for each racial track does seem causally remote from any real-world effects: there are likely to be myriad intervening causes between that decision and the real world effects on particular defendants. However, when the algorithm is *deployed*, classifying the defendant as Black or white determines what other specific features will be brought to bear in determining their risk score. And the effect of *that* racial classification on the defendant is quite direct. Indeed, it is hard to imagine how an effect could be *more* direct than this.

To illustrate, notice that using different racial tracks within an algorithm can lead to scenarios in which classifying a defendant as one race rather than another makes *all the difference* to whether the defendant is determined to be high or low risk. Consider two defendants with an identical criminal history: both defendants have two prior convictions for misdemeanor drug possession. The defendants are also identical with respect to every other feature (e.g., educational history) that a criminal risk assessment algorithm considers when producing risk scores, except one defendant is white and one is Black. Now suppose that each defendant has been found guilty of burglary and is awaiting sentencing. Recidivism risk, calculated by the algorithm, will be one factor considered by the judge in making their sentencing decision. And suppose that, due to concerns about the differential predictive value of criminal history for Black and white defendants, the risk assessment system uses a different racial track for Black and white defendants. For the sake of simplicity, suppose that the only difference between the "white track" and the "Black track" is that the white track adds two points per conviction for drug possession to a defendant's risk score, and the Black track adds one point per conviction for drug possession (to account for the fact that police officers disproportionately target Black Americans for drug arrests). The other features shared by the

two defendants combine to add six more points to each defendant's risk score. The threshold for being designated high risk of recidivism by the system is 10 points. In this scenario, the white defendant will be designated high risk, and the Black defendant will not. Moreover, the only difference between the two defendants that explains this result is the racial classification that led to the white defendant being placed on the white track and the Black defendant being placed on the Black track. If the effect of this racial classification is not direct—in a case where it makes all the difference to the outcome for a defendant—it is hard to imagine that *any* effect would count as direct. Therefore, racial classifications involved in implementing different racial tracks within algorithms can have direct effects on defendants, at least in principle.

It is worth emphasizing here how different the use of different racial tracks is in this respect from the Census. It is extremely difficult to imagine a case where the relevant racial classifications used in the Census make all the difference to specific outcomes for identifiable individuals. The results of the census could eventually impact the proportion of representatives allotted for a given area or the allocation of other government resources, but the effects on specific individuals are not traceable counterfactually to any particular racial classification. This is akin to the way that any election outcome is not traceable counterfactually to any particular vote. By contrast, in the racial tracks case, the effect of being classified as Black on a defendant's risk score is quite easy to trace to the racial classification: we can simply ask whether the same outcome would have eventuated had the defendant been classified as white. In this case the answer is decisively 'No'. In the census case, a counterfactual analysis reveals no decisive answer.

It might be objected that we have misconstrued what makes the effect of a racial classification direct rather than remote. The effect of a racial classification is not direct just because it makes the difference between being labeled high and low risk—that just means the classification

has an impact. Remote effects are still effects.[7] So, if making all the difference between a high and low risk classification is not sufficient for an effect of racial classification to count as direct, what is sufficient? If, as Hellman claims, the use of different racial tracks within an algorithm produces a merely remote effect on identifiable defendants even when it makes all the difference between being labeled high risk or low risk, this is presumably because the point at which the racial classification determines an individual's track is too far causally upstream from the final algorithmic risk classification to constitute a direct effect—that is, the racial classification occurs early on in the risk assessment process to determine "what other factors to employ in making a prediction about recidivism risk" (Hellman 2020, 862). These other factors causally intervene in such a way as to make the racial classification's effect on the defendant causally remote. This further suggests that if the use of different cut points for different racial groups is to have a *direct* effect on a defendant, it must be because it occurs later in the algorithmic risk assessment process such that there are very few intervening causal factors.

We are not in a position to say whether moving the racial classification to the end of the algorithmic risk assessment process could make a legal difference, but it does not appear to constitute a moral difference. To see this, imagine a second way that an algorithmic system might make use of different racial tracks. Suppose that instead of the racial classification occurring early in the algorithmic risk assessment process, racial classification occurred at the end of the process, as a final stage before the system generates a risk score. How would this work? Instead of putting defendants on different tracks within the algorithm as a first step to determine what factors like housing instability or conviction history to consider, the algorithmic system would consider the same factors for each defendant and then adjust each defendant's risk score as a final step, where this

[7] We thank an anonymous reviewer for American Philosophical Quarterly for raising this concern.

adjustment is dependent on the race-relative predictive value of each factor. Suppose that this procedure yields the same risk scores for each defendant as would putting defendants on different racial tracks as a first step. In this case, however, the effects of the racial classification would be direct, as there would be no intervening causal factors between the racial classification and the final risk score for each defendant. Surely whether defendants are placed on different tracks as a first or final step in the algorithmic process cannot make a moral difference to whether the use of race is permissible or not. And therefore, the mere fact that the use of different cut points for different racial groups occurs at the *end* of the algorithmic risk assessment process cannot make for a moral difference between that practice and the use of different racial tracks.

2.4 Different racial tracks within algorithms: No racial generalization?

As Hellman points out, however, not all racial classifications with a direct effect on individuals count as disparate treatment. For example, the use of suspect descriptions involves making racial classifications, and those classifications have a direct effect on suspects; and yet the practice does not constitute disparate treatment. In fact, the use of suspect descriptions is for the most part considered to be legally and morally unproblematic. As long as the use of different racial tracks within algorithms does not make use of racial generalizations, then it does not constitute disparate treatment, even if it has a direct effect on individuals.

Hellman argues that using different racial tracks within algorithms is like the use of suspect descriptions involving race in that neither practice makes use of racial generalizations. She writes:

> [T]he generalization embodied in the algorithm is a generalization about the relationship between housing stability and recidivism, given a person of a particular race. This is analogous to the generalization about the reliability of eyewitness testimony, given a report

> about a perpetrator's race. While the algorithm relies on a generalization about what housing stability or instability indicates for people of each race, the generalization itself is not a racial generalization. It refers to the racial classification but not by relying on a racial generalization (Hellman 2020, 862).

Key to Hellman's argument is that there is *no* generalization about members of a particular race involved in the use of different racial tracks within an algorithm. The only generalization involved in the decision to place a defendant on one track or another is about the reliability of certain types of information given some fact about a person's race.

But it is not clear that this claim can be sustained. To understand why, it will be helpful to get clear about the form of the two types of generalizations involving racial classifications that Hellman has in mind. On the one hand, there are generalizations of the form: Evidence E predicts with probability P trait T for members of group G. Call this kind of generalization an *evidential generalization*. Evidential generalizations are about the reliability of evidence, and they merely refer to a racial classification. On the other hand, *racial generalizations* have the form: Members of (racial) group G possess trait T with probability P. Racial generalizations are not about the reliability of the evidence, but rather, about the members of a racial group themselves. According to Hellman, decisions based upon evidential generalizations do not constitute disparate treatment, while decisions that rely on racial generalizations do.

And yet when an evidential generalization refers to a racial classification and is used to make inferences about specific individuals, this inference can implicitly rely on a racial generalization. And there is good reason to think that the *specific* inference involved in the use of different racial tracks relies on a racial generalization. To see this, let's look more closely at the evidential generalization relied upon in the use of different racial tracks within an algorithm. As Hellman describes it, the

generalization employed by the algorithm is a "generalization about the relationship between housing stability and recidivism, given a person of a particular race."

The problem for the suggestion that this evidential generalization is not about members of any racial group becomes clear if we keep in mind what the generalization is being used to do: make inferences about individuals' probability of reoffending. In order to be useful for predicting recidivism, the generalization about (a) the relationship between housing stability and recidivism, given a person of a particular race, must tell us something about (b) a particular defendant's probability of reoffending. But an inference from (a) to (b) is justified only if we affirm a further generalization that involves attributing a particular property to defendants that fit a certain profile. To make this inference the generalization we must affirm must have roughly this form: A member of (racial) group G possesses trait T1 (e.g., is a risk of reoffending) with probability P, if they possess some further trait T2 (e.g., housing instability). Something like this kind of generalization must bridge the inferential gap between (a) and (b) in any algorithmic risk assessment system that takes different features into consideration depending on the defendant's race.

The crucial point to note, however, is that the generalization that bridges the gap from (a) to (b) is fundamentally a racial generalization. It is about the probability that members of a particular racial group will recidivate. Furthermore, any algorithm that makes use of different tracks for different racial groups will involve a racial generalization of this sort. The use of suspect descriptions involving race does not possess this feature. No further racial generalization (i.e., no generalization about members of some racial group) is required to infer from a particular suspect description that some person is the suspect the police are looking for. There is, then, an important distinction between these two types of inference.

If this is correct, the conditions specified by Hellman under which the use of racial classifications constitutes disparate treatment cannot be used to show that different racial tracks

within an algorithmic system does not constitute disparate treatment. The use of different racial tracks appears both to have direct effects on identifiable individuals and to use racial generalizations. It therefore satisfies the jointly sufficient conditions for disparate treatment.[8]

3. The difference between racial tracks and cut points

We have argued that the use of different racial tracks satisfies the conditions specified by Hellman under which the use of racial classifications constitutes disparate treatment. Thus, Hellman's argument cannot establish a legally relevant distinction between using different racial tracks and different cut points for Black and white defendants. For reasons we gave earlier, if Hellman's argument does not establish a legally relevant distinction between racial tracks and cut points, this is some evidence that it cannot establish a morally relevant distinction either.

However, there is in fact a moral distinction between these two approaches for incorporating race into risk recidivism algorithms—or so we will argue in this section. To preview the basic idea: we propose that the use of different cut points for white and Black defendants is to the disadvantage of all white defendants in a way that the use of different racial tracks is not.[9] After laying out this distinction in greater detail, we explore the extent to which this distinction matters, all things considered, for the moral permissibility including race in algorithmic risk assessment.

To begin, let us assume that employing a higher cut point for Black defendants achieves error parity among Black and white defendants. This is, after all, the promise of applying different cut points in the first place. Assume further that the cut point for white defendants remains

[8] We leave open the possibility that there are other ways of giving algorithmic systems access to race that do not satisfy the conditions for disparate treatment, but we will not address those here.

[9] To clarify, we will speak primarily in terms of *advantages* and *disadvantages* in what follows; at times we will speak of what defendants would *prefer*, or what they might *reasonably reject*. While each term denotes a distinct concept, we find them all to gesture toward the same basic distinction we are interested in, and thus, that they serve roughly the same purpose in this part of the argument.

unchanged. Now let us limit our focus to the class of defendants who will not go on to reoffend upon release. When a higher cut point is employed for Black defendants, the ex ante probability of any given Black defendant being misclassified as high risk is lower than it would have been had they been assessed according to the original cut point. Here "ex ante probability of misclassification" refers to the probability of misclassification prior to the algorithm considering the defendant's recidivism-relevant features. In contrast, when a higher cut point is employed for Black defendants only, white defendants' ex ante probability of misclassification is now *greater than it would have been* had they been judged according to the (new) cut point used for Black defendants. That is, despite achieving error parity, the deployment of different cut points generates what we might call a *counterfactual comparative disadvantage* for white defendants relative to their Black counterparts. Every white defendant—regardless of their risk profiles (e.g. criminal histories, etc.)—would rationally prefer to be assessed according to the cut point used for Black defendants, because their ex ante probability of being misclassified as high risk would be lower. On this approach, then, Blacks and whites are subjected to different standards that every white defendant has self-interested reason to reject. It is worth emphasizing again this is a disadvantage that *every* white defendant faces solely in virtue of their racial classification.[10]

Before returning to the use of different racial tracks, it is worth first asking whether merely counterfactual comparative disadvantage could ground a moral complaint by white defendants. Someone might think that while there is a perceived wrong by white defendants when an algorithm employs race-sensitive cutpoints, there is no unfairness in fact, given that white and Black

[10] Note that the term "disadvantage" is used to refer to *counterfactual comparative* disadvantage. So, when we say that the use of different cut points is to the disadvantage of white defendants, what we mean is that white defendants are worse off *than they would have been* had their recidivism risk score been determined using the cut point used for Black defendants. On the other hand, whites are not any worse off than they *were*, in absolute terms, prior to the introduction of a higher cut point for blacks. We return to this last point below.

defendants were situated very differently to begin with with respect to their error rates. But this response might be too dismissive of the legal value of equality before the law. Marcello Di Bello and Collin O'Neil have argued that the admission of profile evidence in criminal trials is morally objectionable because of the moral requirement on lawmakers not to structure the criminal justice system in a way that creates *ex ante* inequalities in the probability of mistaken conviction for innocent members of different groups. Profile evidence is evidence that expresses a statistical correlation between membership in some group and having committed some crime. About profile evidence they write, " the admission of profile evidence would increase the ex ante risk of mistaken conviction for all and only innocent defendants who match a relevant incriminating profile. This increase in risk would, in turn, expose them to a higher risk of mistaken conviction than other innocent defendants, other things being equal" (Di Bello and O'Neil 2020, 164). Ex ante risk here is the risk of mistaken conviction that is borne by innocent defendants in advance of their identities being known by judges or juries and in advance of any evidence be presented at trial. For the reasons we described above, when lawmakers permit the use of race-sensitive cut points in an algorithmic risk assessment system for the purpose of achieving error parity, they also create an inequality in ex ante risks for white defendants who would not go on to reoffend upon release from detention. Other things being equal, all of these white defendants are more likely ex ante than Black defendants to be mistakenly labeled high risk in virtue of including race-sensitive cut points. We will suggest below that this inequality in ex ante risk imposition on innocent white defendants might be justified, all things considered, as a corrective measure to achieve error parity, but this does not mean that there exists no *pro tanto* objection to the unequal protection of the law entailed by the inclusion of race-sensitive cutpoints.

On the other hand, using different racial tracks within an algorithm need not necessarily give rise to counterfactual comparative disadvantage for all white defendants. Suppose that housing

instability has positive predictive value for white but not Black defendants, and suppose that low educational attainment has positive predictive value for Black but not white defendants. So, on the track for white defendants, housing instability adds points to a defendant's score, while low educational attainment does not. And the reverse is true on the track for Black defendants: housing instability does not add points, but low educational attainment does. One aim of using different racial tracks in this way is to make the algorithmic system more accurate for both Black and white defendants at the group level, since each track incorporates only features that have predictive power for members of that racial group. One possible effect of this improvement in accuracy is that the false positive rate is lower for both Black and white defendants in comparison both to what the group's false positive rate would have been had they been judged according to the status quo race-neutral system and to what the group's false positive rate would have been had they been judged by the other race-specific track. When accuracy for all groups is improved with the result that the false positive rate for each group falls, innocent members of each group now face a lower probability on average of being mistakenly classified as high risk. This is, in a clear sense, to the advantage of both groups.

Suppose, further, that one result of this improvement in accuracy is that false positive rates for both whites and Blacks fall but the algorithm's false positive rate for Blacks falls below the false positive rate for whites. That is, the situation of whites and Blacks is now reversed from what it was prior to the introduction of different racial tracks. Even though there is a new inequality in false positive rates such that the false positive rate for whites is higher than the false positive rate for Blacks, the use of different racial tracks is still *not* to the counterfactual comparative disadvantage of white defendants in the way that implementing different cut points is. Whether being placed on a race-specific track benefits or burdens a particular defendant will depend on what recidivism-risk relevant features that defendant possesses and how those features are taken into account by the

track on which they are placed. It is therefore not true for *every* white defendant who will not go on to reoffend upon release that their ex ante probability of being misclassified as high risk is *greater* than it would have been had they been judged according to the track used for Black defendants. Whether being placed on the white track is to the disadvantage of a given white defendant depends on the recidivism-risk relevant features they possess. This disadvantage cannot be known ex ante, and not all white defendants will face it. Some white defendants would rationally prefer ex post to have been placed on the Black defendants' track. For example, imagine a white defendant who shares all the same recidivism-relevant features as his Black counterpart; however, due to his race, the white defendant has points added to his risk score on the basis of his history of housing instability that would not have been added were he placed on the Black track. On the racial tracks approach, the white defendant faces a higher probability of being classified as high risk and thus being subjected to a longer period of detention than his Black counterpart; his being white could make all the difference between substantially different forms of treatment. But, while this might happen to *some* defendants, it is not true that *every* defendant of one racial group faces a higher probability ex ante of being classified as high risk solely in virtue of being placed on their race-specific track. To highlight this difference between cut points and tracks, notice that being placed on the Black track will be to the disadvantage of some Black defendants as well, depending on their recidivism-relevant characteristics. So, even though *some* white defendants would prefer to be placed on the Black defendants' track, given the manner in which their recidivism-risk relevant features are taken into account by the algorithm, this would not be true of *all* white defendants.

The central morally salient difference between the use of different cut points and the use of different racial tracks, then, is that the use of different cut points produces a clear disparity between the groups: every white defendant will be held to a higher legal standard by virtue of their race than every Black defendant. By contrast, different racial tracks can be implemented so that they are to the

advantage of the average member of both groups, even if particular members of both groups will be worse off than they would have been had the algorithm not used different racial tracks. As we can see, this account of the moral distinction between these two approaches is importantly different from Hellman's legal account of the two. The distinction lies fundamentally in the way each approach affects the ex ante probability of misclassification for members of either group. Because using different racial tracks lowers the ex ante probability of misclassification for the average member of both groups, it is easier to justify: the rationale for deploying separate tracks is not only that it achieves error parity, but that it is to the overall benefit of all groups. It is much harder, at first glance, to justify using different cut points. The core problem, as we saw, is that every white defendant would rationally prefer to be judged according to the cut point used for Black defendants. And ordinarily, deliberately introducing this sort of disparate treatment would constitute a significant barrier to justification.

But merely identifying this moral distinction does not yet establish that the practice of using different racial tracks is morally permissible, while using racial cut points for different racial groups is prohibited. It is important to remember that changing the cut point only for Black defendants does not make white defendants worse off than they were before the introduction of a new cut point for blacks. To the extent that whites could be said to be disadvantaged at all, this is only relative to the standards that apply to Black defendants—that is, in a counterfactual comparative sense. Given that the different cut points are not arbitrarily introduced, but aimed at producing error parity between the two groups, there is a sense in which white defendants are not disadvantaged after all. If one is persuaded that error parity is an important metric of algorithmic fairness, and this approach achieves error parity by introducing different cut points, then there is at least one meaningful sense in which this approach makes matters fairer overall. The use of different cut points achieves error parity while

also being Pareto superior to the status quo ante: white defendants are no better or worse off than they were before, while all Black defendants are much better off.

Moreover, insofar as using different cut points aims at rectifying past systemic injustices that have historically fallen on Black individuals, it may rightly be viewed as a tool for repairing these injustices. It is not uncommon elsewhere in society to implement different standards for different racial groups as a way of rectifying past injustices. This is, after all, a popular moral justification of certain affirmative action programs in hiring or university admissions. For example, in an effort to respond to a history of injustice, in which Black defendants were deprived equal opportunity to join the workforce, companies and industries have instituted preferential hiring practices in which Black applicants are given some preference over their white counterparts. This was also prevalent in university admissions, which eventually led to the U.S. Supreme Court case *Regents of the University of California v. Bakke* (1978), in which Allen Bakke, a white man, argued that was denied entry into medical school on the basis of his race, since his scores were higher than all of the applicants selected as part of the university's policy to reserve 16 spots for 'qualified minorities'. Ultimately, the Court agreed with Bakke (Regents of the University of California v. Bakke 1978).

At least superficially, affirmative action policies look quite similar to the use of different cut points in risk scoring: Black applicants are judged according to a standard that is different from that of white applicants, such that certain Black applicants may receive job offers (or admission to universities) for which white applicants would have been deemed ineligible or underqualified. And this comparison to affirmative action might not be entirely flattering. The use of different cut points might face similar objections as those faced by affirmative action in college admissions and hiring.[11]

[11] The idea of "algorithmic affirmative action" of various kinds has received attention in the recent literature. See: (Chander 2017; Barocas and Selbst 2016, 714–15; Bornstein 2018; Bent 2020; Humerick 2020)

Specifically, much of the philosophical (as opposed to distinctly legal) criticism of affirmative action has focused on the issue of introducing costs to white applicants in order to achieve the relevant gains for Black applicants. On affirmative action in employment, Alan Goldman writes that:

> [W]hat is positive, what works in favor of members of certain groups, is at the same time negative, for it works to exclude members of other groups. Increasing the percentage of nonwhite males will decrease the percentage of white males, and this means in a situation of scarcity that certain white males will be denied jobs they might otherwise have secured. (Goldman 1976, 182)

Even those who are ultimately sympathetic to affirmative action note the problem of costs. For example, Judith Thomson writes that, "choosing this way of making amends means that the costs are imposed on the young white male applicants who are turned away," though she goes on to say that "it is not entirely inappropriate that those applicants should pay the costs" (Thomson 1973, 383). Other scholars have discussed the issue of the distribution of such costs at length (Amdur 1979; Groarke 1990). The basic moral concern is that affirmative action policies are essentially zero-sum: in order to achieve gains for one group, another group must bear the costs. While some find these costs justified, others reject affirmative action policies for this very reason.

Unlike affirmative action, the use of different cut points in recidivism risk scoring, in the way proposed, is not zero-sum for defendants being scored: it makes things better for Black defendants—specifically, by shrinking the positive error rate—without burdening white defendants with any additional cost. Unlike in hiring or admissions, there is not a finite number of risk classifications that all defendants are competing for: one defendant being classified as low-risk does not affect any other defendant's prospects. By lowering the percentage of low risk Black defendants

who are misclassified as high risk, the system does not thereby increase the percentage of low risk white defendants who are mistakenly classified as high risk. It is true, as we argue above, that white defendants would be subjected to a different legal standard. And while white defendants would no doubt prefer to be assessed according to the standard used for Black defendants, their treatment would be no worse by virtue of Black defendants being assessed by a different standard. Thus, unlike affirmative action policies, the use of different cut points does not introduce new costs to some in achieving benefits for others.[12]

Thus, one argument for thinking that the use of cut points is impermissible—one that draws an analogy with affirmative action—fails. Are other arguments more likely to succeed? Perhaps the use of different racial tracks is morally superior to the use of different cut points, since it confers benefits on both groups as opposed to only Black defendants. And thus, if given the choice between one or the other option, but not both, courts ought to opt for the use of different racial tracks.

4. Conclusion

This paper has explored two proposals for including race in recidivism risk algorithms—namely, the use of different racial tracks and different racial cut points. Is there a moral difference between these two methods? We explored Deborah Hellman's legal account of these approaches, the use of different racial tracks is permitted because it does not constitute disparate treatment. Specifically, Hellman's argument implies that the use of different racial tracks–but not the use of different cut points–can avoid the charge of disparate treatment because the effects on individuals are indirect and the racial classification embodied by the algorithm does not rely on a racial generalization. But

[12] Clinton Castro raises a related point for the broader context of machine bias. See: (Castro 2019).

as we have argued, the use of different racial tracks seems to have direct effects and seems to rely on a racial generalization, thereby dissolving the apparent distinction between the two approaches. We then argued, in section 3, that while Hellman's grounds for a distinction between the two approaches is mistaken, there is indeed an important distinction between them. The use of different cut points is to the counterfactual comparative disadvantage, ex ante, of all white defendants, while the use of different racial tracks can in principle be to the advantage of all groups, though some defendants in both groups will fare worse. We then asked whether this moral distinction entails that the use of cut points is impermissible. We conclude that while there is indeed a morally important distinction between these two approaches for incorporating race into recidivism algorithms, it remains an open question whether this distinction makes a difference to their moral permissibility.[13]

Works Cited:

National Institute of Justice. 2021. "2021 Review and Revalidation of the First Step Act Risk Assessment Tool"

Amdur, Robert. 1979. "Compensatory Justice The Question of Costs." *Political Theory* 7 (2): 229–44. https://doi.org/10.1177/009059177900700205.

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. "Machine Bias." *ProPublica*, May 23, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Barocas, Salon, and Andrew D. Selbst. 2016. "Big Data's Disparate Impact." *California Law Review* 104 (3): 671–732.

Bent, Jason. 2020. "Is Algorithmic Affirmative Action Legal?" *Georgetown Law Journal* 108: 803–53.

Bornstein, Stephanie. 2018. "Antidiscriminatory Algorithms." *Alabama Law Review* 70 (2): 519–73.

Castro, Clinton. 2019. "What's Wrong with Machine Bias." *Ergo, an Open Access Journal of Philosophy* 6. https://doi.org/10.3998/ergo.12405314.0006.015.

[13] This work is supported by the National Science Foundation under Grant No. 1917707, and the UF College of Journalism and Communications Consortium for Trust in Media and Technology. We thank Clinton Castro, audiences at UC-Boulder, York University, Florida International University, and reviewers for this journal for their many helpful comments.

Chander, Anupam. 2017. “The Racist Algorithm?” *Michigan Law Review* 115 (6): 1023–45.
Corbett-Davies, Sam, Emma Pierson, Avi Feller, and Sharad Goel. 2016. “A Computer Program Used for Bail and Sentencing Decisions Was Labeled Biased against Blacks. It’s Actually Not That Clear.” *Washington Post*, October 17, 2016. https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/.
Corbett-Davies, Sam, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. “Algorithmic Decision Making and the Cost of Fairness.” In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 797–806. Halifax NS Canada: ACM. https://doi.org/10.1145/3097983.3098095.
Di Bello, Marcello, and Collin O’Neil. 2020. “Profile Evidence, Fairness, and the Risks of Mistaken Convictions.” *Ethics* 130 (2): 147–78.
Equivant. 2018. “Response to ProPublica: Demonstrating Accuracy Equity and Predictive Parity.” Equivant. December 1, 2018. https://www.equivant.com/response-to-propublica-demonstrating-accuracy-equity-and-predictive-parity/.
Flores, Anthony, Kristin Bechtel, and Christopher Lowenkamp. 2016. “False Positives, False Negatives, and False Analyses: A Rejoinder to ‘Machine Bias: There’s Software Used Across the Country to Predict Future Criminals. And It’s Biased Against Blacks.’” *Federal Probation* 80 (2): 38–46.
Goldman, Alan H. 1976. “Affirmative Action.” *Philosophy & Public Affairs* 5 (2): 178–95.
Groarke, Leo. 1990. “Affirmative Action as a Form of Restitution.” *Journal of Business Ethics* 9 (3): 207–13. https://doi.org/10.1007/BF00382646.
Heaven, Will Douglas. 2020. “Predictive Policing Algorithms Are Racist. They Need to Be Dismantled.” *MIT Technology Review*, July 17, 2020. https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/.
Hedden, Brian. 2021. “On Statistical Criteria of Algorithmic Fairness.” *Philosophy and Public Affairs* 49 (2): 209–31.
Hellman, Deborah. 2020. “Measuring Algorithmic Fairness.” *Virginia Law Review* 106 (4): 811–66.
Humerick, Jacob. 2020. “Reprogramming Fairness: Affirmative Action in Algorithmic Criminal Sentencing.” Columbia Human Rights Law Review. April 15, 2020. http://hrlr.law.columbia.edu/hrlr-online/reprogramming-fairness-affirmative-action-in-algorithmic-criminal-sentencing/#post-1397-footnote-ref-118.
Huq, Aziz. 2019. “Racial Equity in Algorithmic Criminal Justice.” *Duke Law Journal* 68: 1043–1134.
Morales v. Daley. 2000, 116 801. S.D. Tex.
Regents of the University of California v. Bakke. 1978. United States Supreme Court.
Skeem, Jennifer, and Christopher Lowenkamp. 2020. “Using Algorithms to Address Trade-offs Inherent in Predicting Recidivism.” *Behavioral Sciences and the Law* 38 (3): 259–78.
Thomson, Judith Jarvis. 1973. “Preferential Hiring.” *Philosophy & Public Affairs* 2 (4): 364–84.