Learning to Discriminate: The Perfect Proxy Problem in Artificially Intelligent Criminal Sentencing

BENJAMIN DAVIES & THOMAS DOUGLAS[1]

[Draft 17 Nov 2020. Commissioned for Roberts J, Ryberg J (eds) *Principled Sentencing and Artificial Intelligence*.]

Abstract. It is often thought that traditional recidivism prediction tools used in criminal sentencing, though biased in many ways, can straightforwardly avoid one particularly pernicious type of bias: direct racial discrimination. They can avoid this by excluding race from the list of variables employed to predict recidivism. A similar approach could be taken to the design of newer, machine learning-based (ML) tools for predicting recidivism: information about race could be withheld from the ML tool during its training phase, ensuring that the resulting predictive model does not use race as an explicit predictor. However, if race is correlated with measured recidivism in the training data, the ML tool may 'learn' a perfect proxy for race. If such a proxy is found, the exclusion of race would do nothing to weaken the correlation between risk (mis)classifications and race. Is this a problem? We argue that, on some explanations of the wrongness of discrimination, it is. On these explanations, the use of an ML tool that perfectly proxies race would (likely) be more wrong than the use of a traditional tool that imperfectly proxies race. Indeed, on some views, use of a perfect proxy for race is plausibly as wrong as explicit racial profiling. We end by drawing out four implications of our arguments.

Keywords. Discrimination; Profiling; Machine Learning; Algorithmic Fairness; Racial Bias; Redundant Encoding; Criminal Recidivism; Crime Prediction.

1. Introduction

Traditional tools for predicting recidivism—often called actuarial risk assessment instruments—employ a fixed number of human-selected variables and a regression-based algorithm to classify the risk that an individual will re-offend. A commonly used example is the Violence Risk Appraisal Guide (VRAG), which employs 12 variables, including age, history of alcohol abuse, and marital status, to classify offenders into one of nine risk categories for violent recidivism.

Recently, there has been much interest in, and work to develop, more sophisticated machine learning-based recidivism prediction tools (Berk and Hyatt 2015). Based on a large set of 'training data' about a population of offenders, including information

about who went on to recidivate,[2] a data mining algorithm would derive a model that can be deployed to assess recidivism risk in other populations.

Recidivism prediction tools, whether of the traditional variety ('traditional tools') or developed through machine learning ('ML tools'), are commonly criticised for being biased against members of certain racial groups. In a well-known example, *ProPublica* criticised the COMPAS algorithm for wrongly predicting recidivism much more commonly in Black Americans than White Americans (Angwin et al. 2016; see also, Angwin and Larson 2016; Chouldechova 2017; Dieterich et al. 2016). In fact, recidivism prediction tools exhibit many different kinds of bias (Barocas and Selbst 2016; Berk et al. Forthcoming; Chouldechova 2017; Hacker 2018; Zehlike et al. 2020), not all of which can be simultaneously avoided (Berk et al. Forthcoming; Chouldechova 2017; Corbett-Davies et al. 2017; Kleinberg et al. 2017). However, it is often assumed that they can straightforwardly avoid one important kind of bias: direct discrimination on the basis of race.[3]

Overview of the Chapter

In this chapter, we explore the justifiability and significance of this assumption, with reference specifically to ML tools. We first (§2) describe how traditional tools can be designed to avoid direct racial discrimination. We then (§3) identify and describe a problem for attempts to extend this strategy to ML tools: though designers of ML tools may, strictly speaking, be able to avoid direct racial discrimination, it is not clear that they can avoid its *wrongness*. In the subsequent three sections, we pursue this thought by distinguishing various explanations for the wrongness of discrimination, in each case asking what the explanation implies for the use of ML tools. These explanations advert to procedural unfairness (§4), bad outcomes (§5) and disrespect (§6). In §7, we conclude, and draw out some practical implications of our argument.

2. Direct Racial Discrimination and Traditional Tools

We will use the term 'direct discrimination' to refer only to direct *racial* discrimination, and we will take it that A engages in direct racial discrimination against B when A treats B less favourably than she treats or would treat comparator individual(s) C, (partly) on the basis of B's membership of racial group G.[4] Direct discrimination (sometimes called 'disparate treatment') is typically contrasted with indirect discrimination (or 'disparate impact'). At a rough approximation, indirect racial discrimination refers to treatment that does not constitute direct racial discrimination, but does have a disproportionate negative impact on members of one or more racial groups.

In line with most existing literature, we assume that using a recidivism prediction tool would constitute direct discrimination against members of a racial group if (a) the tool employs membership of that group as a predictor of recidivism, and (b) predicted recidivism is used as a basis for unfavourable treatment. The focus of this volume is on criminal sentencing, and when recidivism prediction tools are used in sentencing, predictions often *are* used as a basis for unfavourable treatment: those classified into high-risk categories are subjected to longer or otherwise harsher sentences. So (b) is often satisfied, and we will henceforth simply assume it to be satisfied. However, (a) is not normally satisfied by traditional tools, which typically do not employ race as a predictive variable (Starr 2014, 811-2, 824).[5] Thus, the use of traditional tools is typically not directly discriminatory—at least, not by virtue of employing race as a predictor within the tool.

The use of traditional tools might still constitute direct discrimination for reasons extrinsic to the tool.[6] For example, a policymaker might decide to employ a particular tool in part because she is indifferent to the harms it will impose on a particular racial group. Moreover, a traditional tool might be objectionable because it *reflects*—and perhaps *amplifies*—prior direct discrimination,[7] for example, because the data on recidivism used to produce the tool was the product of directly discriminatory policing or juridical practices (e.g. Lum and Isaac 2016), or because past direct discrimination causally contributes to recivdivism in its victims and this is captured by the data (e.g., Lippert-Rasmussen 2014, 283-300).[8] Finally, use of a traditional tool could be—and we suspect often is—*indirectly* discriminatory, for example, because it has a greater negative impact on already disadvantaged racial groups than others, and lacks any benefit sufficient to justify this unequal impact.[9] However, if membership of G is not used as an explicit predictor, it is in principle possible to use traditional tools without engaging in *direct* discrimination. Moreover, this might be thought a significant result, for direct discrimination is often regarded as involving forms of wrongdoing over and above those present in indirect discrimination.[10]

3. Direct Racial Discrimination and ML Tools

Could designers of ML tools avoid direct discrimination in a similar way? The question might initially seem obtuse. After all, it is straightforward to exclude race from the list of predictors employed by an ML tool. This can be done by withholding information about race from the tool during the training phase.

However, supposing there is a correlation between race and measured recidivism in the training data, an ML tool will, given enough data and in the absence of 'de-biasing',[11] likely 'learn' a proxy for race—a combination of other factors that to some

degree captures the correlation between race and measured recidivism.[12] Traditional tools often contain such proxies too, but what sets ML tools apart is that, given the large amount of data they can be fed, they may learn a *perfect* proxy for race—a combination of variables that *fully* captures the correlation between race and measured recidivism, such that including race over and above this combination would have *no effect* on risk classifications.[13] If a perfect proxy were developed, we might legitimately wonder whether direct discrimination had been avoided, or avoided in any more than name.

There are two distinct questions here. First, would deploying an ML tool that includes a perfect proxy for race *literally be* directly discriminatory by reason of including that proxy? Is treating unfavourably on the basis of a perfect proxy for race just one way of treating unfavourably on the basis of race? This will depend in part on how we interpret 'treating unfavourably on the basis of race'. On one reading, A treats B unfavourably on the basis of race if and only if the *concept* of race figures in the causal process that leads to unfavourable treatment.[14] This need not be the case when A treats B unfavourably because B is picked out by a perfect proxy for race. On another reading, however, the causal role of the *concept* of race is irrelevant. What matters is the causal role of race itself—whether B is treated unfavourably because B is in fact of a certain race.[15] If race *causes* the characteristics that serve as the perfect proxy for race, then use of the proxy will, on this reading, constitute direct discrimination.[16]

Although philosophically interesting, we set aside this metaphysical issue and focus on a second question: even if deploying an ML tool that includes a perfect proxy for race would, strictly speaking, avoid direct discrimination, might it nevertheless be morally wrong[17] for similar reasons and to a similar degree as had it included race and thus been directly discriminatory?

In what follows, we distinguish various explanations commonly given for the wrongness of discrimination, broadly construed, and consider what these imply for the wrongness of using an ML tool that excludes race as an explicit predictor but contains a perfect proxy for race; we call the use of such a tool 'perfect proxy profiling'. Throughout, in assessing the wrongness of perfect proxy profiling we will keep two comparators in view: the use of a traditional tool that excludes race as an explicit predictor but contains an imperfect proxy for race ('imperfect proxy profiling'), and the use of an ML tool that *includes* race as an explicit predictor ('explicit profiling'), and which thus directly discriminates. This is because we are motivated by two questions. First, does the move from imperfect proxy profiling (via traditional tools) to perfect proxy profiling (via ML tools) bring us closer, morally speaking, to direct discrimination? And second, does it bring us all the way?

With respect to the first question, we will argue that, on some explanations of the wrongness of discrimination, perfect proxy profiling is, or is likely to be, more seriously wrong than imperfect proxy profiling;[18] if ML tools were to learn a perfect proxy for race, that would be a problem.[19] With respect to the second, we will suggest that, on some explanations, perfect proxy profiling could, depending on the underlying empirical facts, be as wrong as explicit profiling. This implies that, if we exclude race as an explicit predictor only for it to be replaced by a perfect proxy, we will have done nothing to mitigate that wrongness.[20]

Before turning to these arguments, a preliminary remark concerning our terminology. Our ultimate interest in what follows will be in whether, when and why the *use* of recidivism prediction tools would be discriminatory, and therefore wrong. However, we will sometimes describe the tools themselves as discriminatory (or not).[21] We do not mean thereby to suggest that algorithmic tools are agents, that they can act wrongly, or that they can be discriminatory in any non-derivative sense.[22] When we say that a tool is discriminatory, we mean only that using it to determine sentence harshness would be discriminatory, and in virtue of features of the tool, rather than features of the particular, contingent way in which it is used.

4. Procedural Unfairness

We begin with the view that discrimination is wrong because it is procedurally unfair. Procedural unfairness can be contrasted with substantive unfairness (which we consider in §5): procedural unfairness concerns the process via which goods and ills are allocated to different people, whereas substantive unfairness concerns the resulting pattern of distribution of these goods and ills. If an interviewer discounts a candidate's strengths because of her race, this is procedurally unfair, yet it is consistent with things turning out as they should, for example, because the candidate is rightly hired anyway. Conversely, a selection procedure, such as preferring candidates with better formal qualifications, may seem procedurally fair, yet lead to a pattern of distribution that is substantively unfair, for example, because people who never had the chance to complete their education are excluded from the most attractive jobs.

One view of procedural fairness is Aristotle's instruction to "treat like cases alike" (Gosepath 2007),[23] which is normally taken to imply that fairness requires that any two individuals are treated equally unless there is a *morally relevant* (in the context) difference between them (Halldenius 2018).[24] Several attempts to explain the wrongness of discrimination can be thought of as variants on this view, differing only in which differences they take to be morally relevant.

For example, explicit profiling has been criticised on the ground that it treats some less favourably than others on the basis of facts which they cannot or could not control (Boylan 2008; Gardner 1998). We could think of these critiques as asserting the Aristotelian criterion of fairness conjoined with the view that differences in uncontrollable factors are morally irrelevant. Others suggest that explicit profiling is objectionable because it treats some less favourably than others on the basis of differences in group membership, not individual characteristics (Miller 1999, 169; Shin 2018; Thomas 1992). The underlying thought may be that group-based differences are morally irrelevant.

What do these explanations imply for the wrongness of perfect proxy profiling?

Assume, first, the control explanation: explicit profiling is wrong when and because it treats some less favourably than others based on factors beyond the control of both.[25] How would perfect proxy profiling compare to explicit profiling on this explanation? And how would it compare to imperfect proxy profiling? In most cases all three types of profiling will be wrong. Explicit profiling employs at least one uncontrollable factor as a predictor: race. Neither perfect proxy profiling nor imperfect proxy profiling uses precisely that factor, but both normally employ other factors that are beyond an individual's control, such as age and history of parental offending. It would be possible to design a tool that employed only controllable factors such as marital status and individual history of offending. However, we are not aware of any widely used or advocated tool that takes this approach. In practice, all three types of profiling will employ uncontrollable factors, so, on the control explanation, all three are procedurally unfair. There is, of course, a further question of *how* unfair they are. Perhaps it could be argued that one type of profiling is more procedurally unfair than another if it employs more uncontrollable factors, employs a higher proportion of uncontrollable factors, or overall allows uncontrollable factors to more strongly influence treatment. However, we see no reason to suppose in advance that explicit profiling, perfect proxy profiling and imperfect proxy profiling differ in these ways; whether they do will be contingent on precisely which predictors end up in the predictive model.

Consider now the individuality explanation: explicit profiling is wrong when and because it fails to treat people as individuals. This explanation can be understood in various ways,[26] but in our view, the most plausible understanding is offered by Thomas (1992). Thomas suggests that a problem with explicit profiling is that it employs an unjustifiably coarse predictor (race), ignoring other factors that could and should be used to make finer-grained predictions.[27] Understood thus, the individualist objection seems less powerful against ML tools (whether or not they

explicitly employ race) than it is against traditional recidivism prediction tools. ML tools can be given a very large set of data, and so can make finer-grained predictions than traditional tools. However, the explanation does not clearly distinguish perfect proxy profiling or explicit profiling; both are likely to employ (similarly) fine-grained predictions.

In sum, neither the control explanation nor the individuality explanation clearly establishes that perfect proxy profiling is more wrong than imperfect proxy profiling, or less wrong than explicit profiling.

5. Negative Outcomes

A second type of explanation for the wrongness of discrimination adverts to its unfair, unjust or otherwise disvaluable outcomes. In this section, we first (§5.1) set out some specific versions of this explanation, distinguished by the nature of the outcomes they invoke, before (§5.2) considering what they imply for the wrongness of perfect proxy profiling *vis-à-vis* imperfect proxy profiling and explicit profiling. Throughout, we will, for ease of exposition, present the explanations as appealing to the *badness* of outcomes, though, as indicated above, these explanations sometimes appeal to some more specific disvalue, such as injustice. We will also present the explanations as maintaining that *individual instances* of profiling are wrong when and because they produce bad outcomes, although there are also 'collective' variants of the explanations. Individual acts of profiling that do not cause any bad outcome in isolation may form part of a broader pattern of acts that does cause bad outcomes. In such cases, it may be right to legislate against all such profiling, thus making individual cases wrongful simply because they are rightly legally proscribed (Arneson 2006; Gardner 1998; 2017). Alternatively, it may be that individual instances of profiling can be wrong, even if neither individually harmful nor legally prohibited, by virtue of the role that they play in a wider pattern of practices that produces bad outcomes. For example, perhaps engaging in profiling makes one *complicit* in a wrong committed collectively by all who profile.[28]

5.1 Variants of the Explanation

One outcome-based explanation for the wrongness of discrimination—the proportionality explanation—maintains that it contributes to the under-(over)-representation of some racial groups in the most (dis)advantaged positions in society.[29] For instance, explicit profiling may be wrong when and because, partly as a result of the profiling, Black people make up a larger share of population of offenders classified as 'high risk' than of the population at large.[30]

Sometimes, disproportionality does not seem bad in any way. As Binns (2018a, 1) notes, there is a statistical disproportion in predicted recidivism between men and women, explained partly by the fact that men really are much more likely to recidivate (see also Castro 2019, 408).[31] It is not obvious that there is anything bad about this.

Nevertheless, disproportionality is *often* bad, at least instrumentally.[32] For example, it may contribute to stereotypes that will limit future equality of resources, wellbeing or opportunity. Lever (2017) suggests that profiling may lead to harmful essentialist ideas of race on which certain races have inherent predispositions to certain kinds of crime, while Solanke (2017) takes the central wrong of discrimination to be its stigmatizing effect.[33]

Disproportionality may also be instrumentally disvaluable because it aggravates some past or ongoing group-level injustice (e.g. Yost 2017, 273-82). Suppose that racial group *G* has been subject to systematic oppression in the past, as a result of which *G*-members face stronger incentives to commit crime than others, and thus have offended at higher rates. Suppose further that, as a result of their higher offending rates, members of this group are overrepresented among those deemed to be at high risk of recidivism, and thus among those subjected to the harshest criminal sentences. In this case, the harsher than average treatment meted out to *G*-members plausibly aggravates the injustice of the prior oppression by increasing its harmfulness.

We can think of the proportionality explanation as a specific version of a more general type of explanation: discrimination is wrong when and because it results in (or is part of a wider practice that results in) an unjust, unfair or otherwise undesirable pattern of distribution of goods and ills. Some explanations of this type focus, like the proportionality explanation, on the distribution of goods and ills across *groups.* Others focus on their distribution across *individuals*. For instance, one influential view (Knight 2018; Lippert-Rasmussen 2014) holds that discrimination is wrong when and because it fails to produce the best available pattern of inter-individual distribution, as judged by *desert-weighted prioritarianism* (a theory according to which advantages enjoyed by individuals are more morally valuable (i) the worse off the individual is, and (ii) the more deserving they are).

Another type of explanation appeals to bad outcomes that can readily be understood in non-distributional terms.[34] For example, Hosein (Forthcoming) notes that explicit profiling drives a distrust of the state among members of profiled groups. He further argues that the extent of this distrust may create a situation where individuals from

minority ethnic groups are alienated from their own political society and institutions. Such distrust and alienation are bad outcomes regardless of how they are distributed, though the fact that they are unequally distributed may make them worse.[35]

There are also psychological harms to the individuals who are explicitly profiled. These include the psychological effects on the individual who suffers discrimination. But they also include effects on others, including victims' loved ones and members of the group who expect future discrimination. Psychological harms identified as grounding the wrongness of discrimination include humiliation (Boylan 2008; Bou-Habib 2011), anger and resentment (Alexander 1992; Kennedy 1997), distress (Brown et al. 2000) and fear (Lever 2017; Zack 2015, 59).

5.2 Implications for Perfect Proxy Profiling

How does perfect proxy profiling compare to imperfect proxy profiling and explicit profiling when judged against the various outcome-based explanations surveyed above? The short answer is, 'it depends'—on which specific explanation we consider, the nature of the tool employed, and the context in which it is used. Thus, we will not, in this section, seek to derive any straightforward, general answers to these questions. Instead, we describe four considerations that will be relevant to answering them.

The first consideration is relevant to the comparison between imperfect proxy profiling using traditional tools, on the one hand, and perfect proxy profiling or explicit profiling using ML tools, on the other. Whether a move from traditional tools to ML tools will exacerbate the bad outcomes discussed in the previous subsection depends in part on whether traditional tools under- or over-state the correlation between race and measured recidivism. It is tempting to think that, since traditional tools contain only an imperfect proxy for race, they must understate any correlation. This is incorrect.

Consider the following hypothetical. There is a positive correlation between membership of racial group, *G*, and measured (not necessarily actual) recidivism. However, this is explained by several specific correlations that work in opposing directions, with some weakening the overall correlation, and others strengthening it. For example, perhaps *G*-membership is correlated with economic deprivation, which makes recidivism more likely, but also with strong familial and community relationships, which make recidivism less likely. A traditional tool containing only an imperfect proxy for *G*-membership might over-estimate the positive correlation between *G*-membership and recidivism because, say, the objective, demographic

variables used in the tool capture economic factors well, but don't capture the strength of familial or community relationships. By contrast, an ML tool that either explicitly employed or contained a perfect proxy for *G*-membership would capture both factors. It would thus yield a weaker correlation between *G*-membership and predicted recidivism. In this context, a move from the traditional tool to an ML tool would likely mitigate some of the outcome-based concerns mentioned above. For example, *G*-members would be *less over-represented* among those predicted to re-offend by the ML tool than among those predicted to re-offend by the traditional tool.

In other contexts, however, a move from traditional tools to ML tools could strengthen the correlation between race and predicted recidivism. For instance, suppose that social deprivation causally contributes to measured recidivism and that a correlation between such deprivation and race fully accounts for the correlation between race and measured recidivism. Suppose further that a traditional tool employs an official measure of deprivation as a predictor, but that this measure understates the degree to which deprivation is racialised, say, because it is poor at identifying the types of deprivation that are particularly common in minority racial groups. In this context, the traditional tool will understate the correlation between race and measured recidivism. By contrast, an ML tool that explicitly employs or perfectly proxies race will capture the stronger correlation between race and measured recidivism. It will thus exacerbate some outcome-based objections. For example, it will deliver even more disproportionate results.

A second consideration relevant to our moral comparison is the *claim to advantage* of those who are harmed by the adoption of one type of profiling in preference to another.

Suppose, for example, that a move from imperfect to perfect proxy profiling would tend to disadvantage *G*-members; it would slightly increase the proportion of *G*-members predicted to recidivate, and slightly decrease the proportion of others predicted to recidivate. And suppose that desert-weighted prioritarianism is the correct distributive theory: a person's claim to advantage is stronger the worse off she is, and the more deserving she is. Under these assumptions, the move to perfect proxy profiling will likely exacerbate distributive concerns if *G*-members are on average worse off and more deserving than others, and mitigate them if *G*-members are on average better off and less deserving than others.

A third consideration relevant to our moral comparison is the objectivity of the outcomes invoked by the particular outcome-based explanation of the wrongness of discrimination. Some of the bad outcomes surveyed above are objective in the sense that they are wholly independent of anyone's contingent beliefs about or other

attitudes towards the profiling and its consequences. Suppose, for example, that the relevant bad outcome is simply that some racial groups are overrepresented among those predicted to recidivate. Whether a particular form of profiling produces this bad outcome is objective in the sense we have just described. Indeed, the outcome could occur even if no-one knew or suspected that any profiling had taken place.

Others among the bad outcomes cited by the explanations surveyed above are *subjective*—their occurrence is dependent on contingent beliefs about or other attitudes towards the profiling. Perhaps the most obvious examples are psychological harms. Whether and to what degree a person subjected to profiling is humiliated by that profiling will, for instance, depend not only on the racial distribution of risk classifications but also on what people believe about this distribution and how it is determined. Someone categorised as high risk due to explicit profiling is less likely to feel humiliated if they falsely believe the profiling tool neither explicitly employed nor tracked race.

This distinction is particularly relevant to the comparison between perfect proxy profiling and explicit profiling. To our knowledge, all of the objections to profiling that invoke bad *objective* outcomes invoke outcomes that are mediated wholly by the way in which risk classifications (or misclassifications) are distributed across racial groups. Yet perfect proxy profiling and explicit profiling, by definition, produce the same (mis)classifications. So they will fare equally when judged against these objections.

The same conclusion cannot, however, be drawn with respect to objections that invoke bad *subjective* outcomes. Though perfect proxy profiling and explicit racial profiling produce the same distributions of risk (mis)classifications across racial groups, they might nevertheless be perceived differently and thus, for example, give rise to different psychological harms.

The fourth and final consideration is the *publicity* of the profiling. This is relevant only to explanations which invoke subjective outcomes. As noted above, if people do not know that profiling takes place, they are presumably less likely to be humiliated by it than if the profiling is known. Similarly, if people do not know that profiling takes place, this will likely weaken the tendency of profiling to promote widespread beliefs about the link between race and crime, so stigmatization will be less likely.

There are thus some reasons to think that a move from imperfect proxy profiling via traditional tools to perfect proxy profiling via ML tools might mitigate some of the bad outcomes mentioned above. While any profiling can be done secretly,[36] the

workings of complex ML-derived algorithms are less accessible to the average person than traditional tools employing a small number of predictive variables, such as VRAG. Thus, it might be that an ML-derived algorithm learned a perfect proxy for race, and made sentencing recommendations on this basis, without many people knowing about it.

To the extent that discrimination is wrong by virtue of its subjective outcomes, then if an ML tool developed a perfect proxy for race unbeknownst to most people *and* this meant that some bad outcomes were avoided, the profiling would *in one respect* be less wrong than a comparable, publicly open method of profiling. However, it may still be more wrong overall, since profiling people in ways that cannot be subjected to reasonable public scrutiny contravenes a purported publicity requirement for democratic states (Rawls 1996, 68–9).[37]

6. Disrespect

A final influential family of explanations for the wrongness of discrimination centres on respect (Beeghly 2018). Some of these explanations appeal to what we call 'mental state disrespect': discrimination is wrong when and because discriminators act on the basis of disrespectful attitudes or fail to act on the basis of respectful attitudes. For instance, discriminators might view members of racial minorities with unjustified hatred (Garcia 1996); inaccurately judge them to have less than equal moral worth (Alexander 1992); be culpably indifferent to their suffering (Alexander 2010, 203); or simply fail to take account of them as their moral worth demands (Eidelson 2015).[38] As noted in the introduction, while such possibilities are important, they do not raise any particular issue for perfect proxy profiling. Such profiling is plausibly no more or less likely to be based on (dis)respectful attitudes than imperfect proxy profiling. It is also, we think, not obvious that (imperfect or perfect) proxy profiling would fare any better than explicit profiling with respect to mental state disrespect. It seems initially plausible that agents motivated by the most disrespectful attitudes would be more likely than agents with better attitudes to employ explicit profiling, since they are less likely to see it as wrong. But there are reasons to doubt this; given strong legal protections against, and social disapproval of, explicit profiling, such agents might in fact be attracted to proxy profiling as a way of masking their disrespectful attitudes (Barocas & Selbst 2016, 692-3).

We will not further pursue the thought that the use of recidivism prediction tools might be disrespectful on a mental state account. Instead, we focus on an objective understanding of disrespect. Accounts which view discrimination as wrong because objectively disrespectful have been influential (Glasgow 2009, 2015; Hellman 2008;

2018; Scanlon 2008; Shin 2018). While such accounts vary in their detail, they share a commitment to the idea that discrimination can be disrespectful, and therefore wrong, because of its *objective social meaning*—for example, because it sends the message that members of a particular racial group are inferior—even if there is no disrespectful attitude (or lack of a respectful attitude) behind it. Actions can have an objective social meaning—where objectivity implies independence from any particular individual's contingent mental states—due to various social and historical facts, including facts about how (reasonable) people typically respond when such acts are performed.

Consider, for instance, purchasing a product in a shop. What it means to hand over pieces of metal, paper, or plastic is not independent of belief altogether, but its meaning does not depend on the *particular* intentions of specific individuals. If Marilyn hands over some legal tender, and a shopkeeper lets her take away some chocolate, Marilyn has *bought* that chocolate regardless of what she believes has happened or intended to happen. Similarly, say proponents of the objective meaning explanation, a discriminatory act can have a disrespectful meaning even if the discriminator has exemplary attitudes, and even if the victim neither knows nor suspects the act to be discriminatory (Glasgow 2015, 121). For example, it may have the meaning because many victims reasonably believe that acts of this type are normally motivated by objectionable attitudes.

Suppose the social meaning account of disrespect is correct.[39] This raises the issue of what the social meaning would be of using an ML tool where (a) the tool contains a perfect proxy for the offender's racial group, and (b) the tool classifies the offender as high risk partly on the basis of this proxy.

One option is to see ML tools as both novel and unique, and hence currently lacking any social meaning. On this view, while the use of ML tools might over time *develop* a social meaning, it currently lacks one. This would mean, *a fortiori*, that perfect proxy profiling cannot be objectively disrespectful.

However, racial profiling and offender risk-assessments based on traditional tools are now familiar practices, and it is plausible to think that the social meaning of these practices would immediately extend to the use of ML tools. The interesting question is whether the meaning of perfect proxy profiling (via use of an ML tool that excludes race as an explicit predictor) would be closer to that of explicit profiling (via use of an ML tool that includes race as an explicit predictor), or to that of imperfect proxy profiling (via use of a traditional tool). We suspect that it would be closer to the former. This can be supported by considering the following imaginary example.

Suppose that there are only two towns—Springfield and Shelbyville—in a sparsely populated region of the USA. The residents of Springfield regard all residents of the closest town, Shelbyville, with hatred, simply because they are from Shelbyville. However, no Springfielder has ever been to Shelbyville, while Shelbyville residents—knowing about their neighbours' attitudes—avoid Springfield. All residents of both towns are thus unaware that all residents of Springfield are White, while all Shelbyville residents are Black. One day Bart, a Springfielder, decides to go to Shelbyville to steal their town monument. On arrival, he realises the facts about the racial divide. He also learns the reasons for the divide: several generations earlier, a combination of explicit racism on the part of White residents, and racialised economic disadvantage suffered by Black residents, caused Black Springfielders to leave and establish Shelbyville. While these facts were not actively hidden in subsequent generations, they were not widely discussed, and so most residents of both towns are unaware of their joint history. When Bart gets back to Springfield, he tells all his friends these facts, and soon most Springfielders know them. Yet their negative attitudes persist.

Both before and after Bart's visit, Springfielders engage in some actions—such as avoiding Shelbyville and expressing hatred for Shelbyvillians—that are motivated in part by their hatred of Shelbyvillians. What should we say about the social meaning of these actions? Before Bart's visit, Springfielders hate Shelbyvillians, but no resident of either town has reason to think that 'Shelbyvillian' is co-extensive with 'all and only the Black residents of my local region', nor that it picks out a group whose forebears were victimised by Springfielders. While actions motivated by this hatred might have a morally questionable social meaning, it is very doubtful that they express racial disrespect. But once Bart and most other Springfielders realise the racial extension of their attitudes, and the underlying historical facts, things plausibly change. Saying "I want nothing to do with anyone from Shelbyville" very plausibly takes on a racially disrespectful meaning now that it is widely known that this covers all and only the Black residents of one's local region, and that Black people have historically been disadvantaged, in part due to explicit racism on the part of one's own predecessors. Moreover, it is plausible that it takes on this meaning even in cases where the particular person making the utterance is one of the few Springfielders who still does not know the relevant racial and historical facts, and even where this ignorance is blameless.

Similarly, our view is that, at least once it becomes clear and widely known that an ML tool has learned a perfect proxy for membership of a racial group known to have been subject to systematic and racially motivated oppression, use of the tool would

plausibly come to express distinctively racial disrespect. At least, we think this would plausibly be so if use of the tool is accompanied by no efforts to mitigate its racialised meaning or disparate racial impacts, and if members of the oppressed racial group reasonably regard use of the tool as manifesting an indifference to the past injustice on the part of those who designed or use the tool (whether or not it in fact manifests such an indifference).[40] Indeed, it seems plausible that use of the tool in this context would be just as disrespectful as would using a tool that explicitly uses membership of the racial group in question, holding fixed the attitudes of those who designed and use the tool.

By contrast, the presence of a racially disrespectful social meaning becomes far less clear when the proxy is imperfect and understates the correlation between race and measured recidivism. Consider a variant of the Springfield-Shelbyville case in which some migration of Black people from Springfield to Shelbyville still occurred, and for the same reasons, but many Black people remained in Springfield, and many White people also migrated to Shelbyville for reasons unconnected to race. As a result, Springfield is only 55% White, and Shelbyville 55% Black. In this version of the case, it is less clear that the Springfielders' hatred-motivated actions have a racist meaning once the facts are widely known. Similarly, it is not clear to us that continuing to use a traditional recidivism prediction tool after discovering that one of the predictors it employs is weakly correlated with race has the same social meaning as continuing to use a tool discovered to contain a perfect proxy for race.

To be clear, we are not suggesting here that use of an imperfect proxy would never express racial disrespect or even that it would always express *less* disrespect than use of a perfect proxy.[41] Rather, our claim is that, at least in some cases—most likely when the proxy is weak—it would express no or less disrespect. This, then, may be one respect in which perfect proxy profiling is more seriously wrong than at least some instances of imperfect proxy profiling.

7. Conclusion and Implications for Sentencing

We have considered three different types of explanation for the wrongness of discrimination, adverting respectively to procedural unfairness, bad outcomes, and disrespect. What do these explanations imply for the wrongness of ML-based perfect proxy profiling in criminal sentencing, *vis-à-vis* ML-based explicit profiling, and imperfect proxy profiling using traditional tools?

One version of the appeal to procedural unfairness holds that discrimination is wrong when and because it treats some less favourably than others on the basis of factors

beyond their control. This explanation does not clearly distinguish perfect proxy profiling from imperfect proxy profiling or explicit profiling; they all employ uncontrollable factors as predictors and there is, we think, no reason to suppose in advance of looking in detail at a particular tool that one type of tool will rely more heavily on such factors than another. Another version holds that discrimination is wrong when and because it fails to treat people as individuals. This explanation again fails to clearly distinguish perfect proxy profiling from explicit profiling. However, it may suggest that both, being ML-based, are typically less wrong than imperfect proxy profiling using traditional tools, since ML tools would normally employ finer-grained predictors.

Consider next explanations that appeal to bad outcomes. It is difficult to draw any straightforward and generalisable conclusions regarding the implications of these explanations for the three types of profiling under consideration. However, one relatively clear implication is that perfect proxy profiling will fare no better or worse than explicit profiling with respect to explanations that invoke bad *objective* outcomes. Another is that, to the extent that imperfect proxy profiling understates the correlation between race and measured recidivism, a move to ML tools that either include race or a perfect proxy for it will tend to strengthen racial disproportionality, increasing the degree to which members of the group picked out by the proxy are overrepresented amongst those assigned the highest risk classifications, and thus amongst those given the harshest sentences. It will thus tend to exacerbate the (further) bad outcomes, such as stigmatization, to which such disproportionality often contributes.

Consider, finally, disrespect-based explanations. None of the three types of profiling under consideration need involve mental state disrespect, and it is an open empirical question which is or are more likely to do so; however, all could involve social meaning disrespect. We suggested that, in certain contexts, perfect proxy profiling would plausibly be more disrespectful, with regard to its social meaning, than imperfect proxy profiling, and indeed that it may be as disrespectful as explicit profiling.

What implications might our arguments have? Let us briefly mention four.

First, our arguments cast doubt on the thought that, as we move from traditional to ML tools, we should retain the practice of excluding race from our predictive models. On none of the explanations that we have surveyed is it *obvious* (though on some it is *arguable*) that perfect proxy profiling using ML tools is, or is typically, less seriously wrong than explicit profiling using such tools. This suggests that, if an ML

tool would indeed learn a perfect proxy for race, preventing it from explicitly employing race may not mitigate its wrongness.

Second, some of the views that we have considered are reassuring about the move from traditional to ML tools, whether or not those latter tools explicitly employ race. If the appeals to procedural fairness and/or mental state disrespect exhaust the wrongness of discrimination, there may be no reason to fear the move from traditional to ML tools, since, on these views, neither perfect proxy profiling nor explicit profiling via ML tools is clearly more seriously wrong than imperfect proxy profiling.

Third, on other views, however, the move to ML tools is of concern. Suppose that explanations appealing to bad outcomes or objective social meanings account for some of the wrongness of discrimination. On these views, imperfect proxy profiling is already morally problematic, but a move to either explicit or perfect proxy profiling exacerbates some of these problems. This suggests that the move to ML tools may bring with it a greater need for de-biasing measures.

Finally, a fourth possible implication of our arguments is that they may cast doubt on one objection to such de-biasing. De-biasing methods typically require that the recidivism prediction tool explicitly take race into account, for example, in order to adjust classifications or segment analysis on the basis of race (e.g., Corbett-Davies and Goel 2018, esp. 2, 14). One possibility, for example, is to set a different risk threshold for deeming a person to be high-risk depending on their racial group.[42] Yet factoring race into tools is sometimes thought problematic, perhaps because it is taken to involve direct discrimination.[43] Hellman (2020) challenges this view at the level of law, arguing that explicitly employing race in an algorithm will not always be unlawful in the US.[44] Our analysis suggests another kind of response. Even if de-biased algorithms would be directly discriminatory, that might not give us a decisive moral reason to eschew them, since 'raw' (i.e. not de-biased) algorithms may be morally wrong in similar ways and to a similar degree. This will be plausible if direct discrimination is wrong by virtue of its disrespectful social meaning, bad objective consequences, procedural unfairness, or some combination of these. We argued that none of these factors clearly distinguish perfect proxy profiling from explicit profiling.

Bibliography

Alexander, Larry. 1992. "What makes wrongful discrimination wrong? Biases, preferences, stereotypes, and proxies". *University of Pennsylvania Law Review 141.* pp. 149–220

Alexander, Michelle. 2010. *The New Jim Crow: Mass Incarceration in the Age of Colorblindness.* New York: The New Press

Angwin, Julia, Jeff Larson, Surya Mattu and Lauren Kirchner. 2016. "Machine Bias". *ProPublica*

Angwin, Julia and Jeff Larson. 2016. "ProPublica Responds to Company's Critique of Machine Bias Story. *ProPublica*. 29 July 2016. https://www.propublica.org/article/propublica-responds-to-companys-critique-of-machine-bias-story [Accessed 31 July 2020].

Arneson, Richard. 2006. "What is wrongful discrimination?" *San Diego Law Review 43* (4): pp. 775–808.

Barocas, Solon and Andrew D. Selbst. 2016. "Big Data's Disparate Impact". *California Law Review 104* (3): pp. 677–690.

Beeghly, Erin . 2015. "What is a stereotype? What is stereotyping?" *Hypatia 30* (4): pp. 675–691

-------2018. "Discrimination and disrespect". In *The Routledge Handbook of the Ethics of Discrimination*, edited by Kasper Lippert-Rasmussen, pp. 83–96. Abingdon: Routledge.

Berk, Richard and Jordan Hyatt. 2015. "Machine Learning Forecasts of Risk to Inform Sentencing Decisions". *Federal Sentencing Reporter 27* (4): pp. 222–28.

Berk, Richard, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Forthcoming. "Fairness in Criminal Justice Risk Assessments: The State of the Art". *Sociological Methods & Research*

Bierria, Alisa. Forthcoming. "Racial conflation: Rethinking agency, black action, and criminal intent". *Journal of Social Philosophy*.

Binns, Reuben. 2018a. "Fairness in machine learning: Lessons from political philosophy". *Proceedings of Machine Learning Research 81*: pp. 1-11

-------2018b. "What can political philosophy teach us about algorithmic fairness?" *IEEE Security and Privacy Magazine 16* (3): pp. 73–80.

Bou-Habib, Paul. 2011. "Racial profiling and background injustice". *Journal of Ethics 15*: 33–46

Boylan, Michael. 2008. "Racial Profiling and Genetic Privacy". *Center for American Progress*

Brown, Tony, David Williams, James Jackson, Harold Neighbors, Myriam Torres, Sherrill Sellers and Kendrick Brown. 2000. "Being black and feeling blue: The mental health consequences of racial discrimination". *Race & Society 2* (2): pp. 117–131.

Calders, Toon and Indrė Žliobaitė. 2013. "Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures". In *Discrimination and Privacy in the Information Society: Data Mining and Profiling in Large Databases* edited by Bart Custers, Toon Calders, Bart Schermer and Tal Zarsky, pp. 43–57. Berlin, Heidelberg: Springer.

Castro, Clinton. 2019. "What's wrong with machine bias?" *Ergo 6* (15): pp. 405–426

Chouldechova, Alexandra. 2017. "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments". *Big Data 5* (2): pp. 153-163.

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* pp. 797–806. New York: Association for Computer Machinery.

Corbett-Davies, Sam and Sharad Goel. 2018. "The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning". ArXiv:1808.00023 [Cs]. http://arxiv.org/abs/1808.00023 [Accessed 31 July 2020].

Crenshaw, Kimberlé. 1989. "Demarginalizing the intersection of race and sex: a black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics." *University of Chicago Legal Forum 1*: pp. 139–167.

Custers, Bart, Toon Calders, Bart Schermer and Tal Zarsky (eds.) 2013. *Discrimination and Privacy in the Information Society: Data Mining and Profiling in Large Databases* Berlin, Heidelberg: Springer.

Dieterich, William, Christina Mendoza and Tim Brennan. 2016. *COMPAS Risk Scales: Demonstrating Accuracy, Equity and Predictive Parity*. Northpointe Inc. https://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf [Accessed 31 July 2020]

Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold and Richard Zemel. 2012., "Fairness through awareness". *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pp. 214–226. Cambridge, Massachusetts: Association for Computing Machinery.

Eidelson, Benjamin. 2015. *Discrimination and Disrespect*. Oxford: OUP

European Union. 2000. "Council Directive 2000/43/EC implementing the principle of equal treatment between persons irrespective of racial or ethnic origin". *Official Journal L 180, 19/07/2000 P. 0022–0026* https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32000L0043&from=FR [Accessed 30 July 2020].

Frase, Richard S. 2014. "Recurring Policy Issues of Guidelines (and Non-Guidelines) Sentencing: Risk Assessments, Criminal History Enhancements, and the Enforcement of Release Conditions". *Federal Sentencing Reporter 26* (3): pp. 145–157.

Garcia, Jorge. 1996. "The heart of racism". *Journal of Social Philosophy 27*: pp. 5–45.

Gardner, John. 1996. "Discrimination as Injustice". *Oxford Journal of Legal Studies 16*: 353–367

-------1998. "On the ground of her sex(uality)". *Oxford Journal of Legal Studies 18*: pp. 167–187

-------2017. "Discrimination: The good, the bad, and the wrongful". *Proceedings of the Aristotelian Society 118*: pp. 55–81.

Glasgow, Joshua. 2009. "Racism as disrespect". *Ethics 120* (1): pp. 64–93

-------2015. "The meaning and wrongness of discrimination". *Criminal Justice Ethics 34*: pp. 116–129.

Gosepath, Stefan. 2007. "Equality". *Stanford Encyclopedia of Philosophy*.

Hacker, Philipp. 2018. "Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies against Algorithmic Discrimination under EU Law". *Common Market Law Review 55* (4): pp. 1143-1186

Halldenius, Lena. 2018. "Discrimination and irrelevance". In *The Routledge Handbook on the Ethics of Discrimination*, edited by Kasper Lippert-Rasmussen, pp. 108–18. Abingdon: Routledge.

Hardt, Moritz, Eric Price and Nathan Srebro. 2016. "Equality of Opportunity in Supervised Learning". In *Advances in Neural Processing Systems 29* edited by Daniel D. Lee, Ulrike von Luxburg, Roman Garnett, Masashi Sugiyama and Isabelle Guyon, pp. 3323-31. Red Hook, NY: Curran Associates

Heinrichs, Bert. 2007. "What is discrimination and when is it morally wrong?" *Jahrbuch für Wissenschaft und Ethik 12* (1): pp.97–114

Hellman, Deborah. 2008. *When is Discrimination Wrong?* Cambridge, MA: Harvard University Press.

-------2018. "Discrimination and social meaning". In *The Routledge Handbook on the Ethics of Discrimination* edited by Kasper Lippert-Rasmussen. pp. 97–107. Abingdon: Routledge.

-------2020. "Measuring Algorithmic Fairness". *Virginia Law Review 106* (4): pp. 811–66.

Hosein, Adam Omar. Forthcoming. "Racial profiling and a reasonable sense of inferior political status". *The Journal of Political Philosophy 26*: pp. e1–e20

Huq, Aziz. 2019. "Racial equity in algorithmic criminal justice". *Duke Law Journal 68* (6): pp. 1043–1134

Kennedy, Randall. 1997. *Race, Crime and the Law*. New York: Vintage Books.

Khaitan, Tarunabh. 2015. "Indirect discrimination in US and UK law". *OUPblog* https://blog.oup.com/2015/07/indirect-discrimination-us-uk-law/. Accessed 2nd October 2020.

-------2018. "Indirect discrimination". In *The Routledge Handbook on the Ethics of Discrimination* edited by Kasper Lippert-Rasmussen, pp. 30–41. Abingdon: Routledge.

Kleinberg, Jon, Sendhil Mullainathan and Manish Raghavan. 2017. "Inherent Trade-Offs in the Fair Determination of Risk Scores". In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)* edited by Christos H. Papadimitrou, pp. 43:1–43:23. Leibniz: LIPIcs

Knight, Carl. 2018. "Discrimination and equality of opportunity". In *The Routledge Handbook on the Ethics of Discrimination* edited by Kasper Lippert-Rasmussen, pp. 140–150. Abingdon: Routledge.

Kroll, Joshua A., Solon Barocas Edward W. Felten, Joel R. Reidenberg, David G. Robinson and Harlan Yu. 2017. "Accountable Algorithms". *University of Pennsylvania Law Review 165* (3): pp. 633–706.

Kutz, Christopher. 2000. *Complicity*. Cambridge: Cambridge University Press

Lammy, David. 2017. *The Lammy Review: An independent review into the treatment of, and outcomes for, Black, Asian and Minority Ethnic individuals in the Criminal Justice System*

https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/643001/lammy-review-final-report.pdf [Accessed 29 July 2020].

Lepri, Bruno, Nuria Oliver, Emmanuel Letouzé, Alex Pentland and Patrick Vinck. 2018. "Fair, Transparent, and Accountable Algorithmic Decision-Making Processes: The Premise, the Proposed Solutions, and the Open Challenges". *Philosophy and Technology 31* (4): pp. 6111–627.

Lever, Annabelle. 2005. "Why racial profiling is hard to justify: A response to Risse and Zeckhauser". *Philosophy & Public Affairs 33* (1): pp. 94–110

-------2017. "Racial profiling and the political philosophy of race". In *The Oxford Handbook of Philosophy and Race* edited by Naomi Zack, pp. 425–435. Oxford: OUP.

Lim, Desiree. 2019. "The indirect gender discrimination of skill-selective immigration policies". *Critical Review of International Social and Political Philosophy 22*(7): pp.906–928.

Lippert-Rasmussen, Kasper. 2008. "Discrimination and the aim of proportional representation". *Politics, Philosophy & Economics 7* (2): pp. 159–82.

-------2011. "'We are all different': Statistical discrimination and the right to be treated as an individual". *Journal of Ethics 15* (1-2): pp. 47–59.

-------2014. *Born Free and Equal? A Philosophical Inquiry Into the Nature of Discrimination* Oxford: OUP

Lum, Kristian and William Isaac. 2016. "To predict and serve?" *Significance 13* (5): pp. 14–19.

Miller, David. 1999. *Principles of Social Justice.* London: Harvard University Press.

Mogensen, Andreas. 2019. "Racial profiling and cumulative injustice". *Philosophy and Phenomenological Research 98* (2): pp. 452–477

Monahan, John. 2014. "The Inclusion of Biological Risk Factors in Violence Risk Assessments. In *Bioprediction, Biomarkers, and Bad Behavior: Scientific, Legal and Ethical Implications*, edited by Ilina Singh, Walter Sinnott-Armstrong and Julian Savulescu, pp. 57–76. New York: Oxford University Press.

O'Neill, Onora. 2006. "Transparency and the ethics of communication". In *Transparency: The Key to Better Governance?* edited by Christopher Hood and David Heald, pp. 75–89. Oxford: OUP.

Rawls, John. 1996. *Political Liberalism*. New York: Columbia University Press

Royal Society. 2017. *Machine learning: The power and promise of computers that learn by example.* https://royalsociety.org/-/media/policy/projects/machine-learning/publications/machine-learning-report.pdf [Accessed 30 July 2020].

Scanlon, Thomas. 2008. *Moral Dimensions: Permissibility, Meaning, and Blame*. Cambridge, MA: Belknap Press

Segall, Shlomi. 2012. "What's so Bad about Discrimination?" *Utilitas 24* (1): pp. 82–100.

Segev, Re'em. 2018. "Discrimination and law enforcement". In *The Routledge Handbook on the Ethics of Discrimination* edited by Kasper Lippert-Rasmussen pp. 324–334 Abingdon: Routledge.

Shelby, Tommie. 2002. "Is racism in the 'heart'?" *Journal of Social Philosophy 33* (3): pp. 411–420

Shin, Patrick. 2018. "Discrimination and race". In *The Routledge Handbook on the Ethics of Discrimination* edited by Kasper Lippert-Rasmussen pp. 196–206. Abingdon: Routledge.

Slobogin, Christopher. 2012. "Risk Assessment and Risk Management in Juvenile Justice". *Criminal Justice 27* (4): pp. 10–25.

Solanke, Iyiola. 2017. *Discrimination as stigma: A theory of anti-discrimination law.* Oxford: Hart Publishing.

Starr, Sonja B. 2014. "Evidence-Based Sentencing and the Scientific Rationalization of Discrimination". *Stanford Law Review 66* (4): pp. 803–72.

Thomas, Laurence. 1992. "Statistical Badness" *Journal of Social Philosophy 23* (1): pp. 30–41

Thomsen, Frej Klem. 2011. "The Art of the Unseen – Three Challenges for Racial Profiling". *The Journal of Ethics 15* (1): pp. 89–117

-----2018a. "Direct Discriminiation". In *The Routledge Handbook on the Ethics of Discrimination* edited by Kasper Lippert-Rasmussen, pp. 19–29. Abingdon: Routledge.

-----2018b. "Concept, Principle and Norm - Equality before the Law Reconsidered". *Legal Theory 24* (2): pp. 103–134

Walton, Anthony. 1989. "Willie Horton and Me". *The New York Times*. 20 August 2020. http://www.nytimes.com/1989/08/20/magazine/willie-horton-and-me.html [Accessed 30 July 2020].

Yost, Benjamin. 2016. "What's wrong with differential punishment?". *Utilitas 29* (3): pp. 257–285.

Zack, Naomi. 2005. "Race and racial discrimination". In *The Oxford Handbook of Practical Ethics* edited by Hugh Follette pp. 245–271. Oxford: OUP

Zehlike, Meike, Philipp Hacker and Emil Wiedermann. 2020. "Matching Code and Law: Achieving Algorithmic Fairness with Optimal Transport". *Data Mining and Knowledge Discovery 34* (1): pp. 163–200.

Zimmermann, Annette, Elena Di Rossi and Hochan Kim. 2020. "Technology can't fix algorithmic injustice". *Boston Review*. 9 January 2020. http://bostonreview.net/science-nature-politics/annette-zimmermann-elena-di-rosa-hochan-kim-technology-cant-fix-algorithmic

---

[1] We thank, for comments on earlier versions of this chapter, Gabriel De Marco, Adam Omar Hosein, Maximilian Kiener, Kasper Lippert-Rasmussen, Frej Klem Thomsen, and the editors of this collection. For their funding, we thank the European Research Council (Consolidator Award 819757).

[2] This data might include, for example, "information about all contacts, of any sort, with police, social services, health services, and child welfare services" and "data from "dragnet surveillance tools," [and] closed-circuit television ("CCTV") cameras used to acquire and track license plate numbers" (Huq 2019, 1061).

[3] Hacker (2018, 1151-2).

[4] Our definition here is inspired by that contained in European law. See, for example, Council Directive 2000/43/EC (2000) Art. 2, 2(a).

[5] See also Berk and Hyatt (2015, 226); Corbett-Davies et al. (2017, 797, 804); Frase (2014, 149); Hellman (2020, 848); Kroll et al. 2017, 682, 685); Monahan (2014); Slobogin (2012, 13-14).

[6] This point is made by Lippert-Rasmussen, this volume.

[7] Here we mention only two mechanisms via which this may occur. For others, see Barocas and Selbst (2016, 692-3); Calders and Žliobaitė (2013).

[8] Use of the tool in this context might, for example, make the tool-user *complicit* in the earlier discrimination.

[9] See, Lippert-Rasmussen, this volume, for discussion.

[10] This is suggested by the fact that direct discrimination is generally regulated separately from—and indeed often more stringently than—indirect discrimination (see, for discussion, Khaitan 2015). The thought underpinning this regulatory separation is, we suspect, that direct discrimination is, other things being equal, more seriously or at least differently wrong than indirect discrimination. We are not ourselves committed to this view, and indeed our subsequent arguments imply that it is false on certain explanations of the wrongness of direct discrimination. However we suspect many would accept it. Indeed, it is sometimes suggested that indirect discrimination is only of concern insofar as it is evidence of direct discrimination (e.g. Lim 2019, 912-3)

[11] By 'de-biasing', we mean alteration of a recidivism prediction tool, or the data fed into it, to improve the distribution of risk (mis)classifications across racial groups. For discussion of particular de-biasing techniques see, for example, Custers et al. (2013, 223-270); Dwork et al (2012); Hardt et al. (2016). For reviews of various strategies, see Berk et al. (Forthcoming, 24-32); Corbett-Davies and Goel (2018, 8); Hacker (2018, 1176-7); Kroll et al. (2017, 682-692); Lepri et al. (2018, 615-18).

[12] Barocas and Selbst (2016, 721, 728); Calders and Žliobaitė (2013, 54); Hardt et al. (2016, 1); Hacker (2018, 1149); Huq (2018, 1053, 1099-1100); Royal Society (2017, 92).

[13] This phenomenon is often referred to as 'redundant encoding'. See, for discussion, Barocas and Selbst (2016, 695); Dwork et al. (2012, 226); Hacker (2018); Zimmermann et al. (2020).

[14] This is what philosophers would call a '*de dicto*' ('about what is said') reading.

[15] This is what philosophers would call a '*de re*' ('about the thing') reading.

[16] See Thomsen (2018a, 24) and Khaitan (2018, 36-37) for discussion of similar distinctions.

[17] Throughout, we use 'wrong' to mean '*pro tanto* wrong', i.e. in one way wrong, unless otherwise specified.
[18] We will sometimes say 'more wrong' in place of 'more seriously wrong'. We understand both to mean 'possesses a greater degree of *pro tanto* wrongness'.
[19] It tends to be assumed in the literature that the development of a perfect proxy for race would be a problem. See, for example, Hacker (2018, 1148-9); Moritz et al. (2016, 1). The main contribution of this chapter is to begin to explain *why* it is a problem.
[20] See, for a similar view, Huq (2019, 1094).
[21] See also Lippert-Rasmussen, this volume.
[22] Although we do consider this possibility in §6.
[23] See also Gardner (1996).
[24] For discussion and criticism of this principle, see Thomsen (2018b).
[25] For critical discussion, see Heinrichs (2007, esp. 105-6).
[26] For critical discussion see Beeghly (2015, 686-8); Castro (2019, 409-11); Eidelson (2015, Ch5); Hellman (2008, 128); Lippert-Rasmussen (2011, 51-2; 2014, 275-8); Mogensen (2019, 456); and Segev (2018).
[27] For critical discussion, see Lippert-Rasmussen (2011, esp. 49-53).
[28] For a study of individual wrongdoing through complicity in collective wrongdoing, see Kutz (2000).
[29] See Lammy (2017) for discussion of disproportionality in the UK's criminal justice system.
[30] For discussion, see Lippert-Rasmussen (2014, Ch. 7).
[31] For an argument to the effect that disproportionality is not bad *in itself*, see Lippert-Rasmussen (2008).
[32] Disproportionality is often also good evidence of other injustices, such as the presence of institutional barriers or biases that disrupt equality of opportunity (Lippert-Rasmussen 2014, 199; Yost 2017, 272-3), or the clustering of various types of disadvantages on members of a group (Castro 2019). In the context of recidivism prediction, disproportion between racial groups in those deemed to be highest risk might be suggestive of biased policing practices, or of the clustering of forms of social deprivation that tend to promote crime.
[33] See also Walton (1989) and Bierria (Forthcoming).
[34] For discussion of a range of such outcomes, see Thomsen (2011, 105-7).
[35] Hosein (Forthcoming, e1) suggests that this alienation may create a *de facto* "inferior political status" (see also Alexander 2010; Crenshaw 1989).
[36] For instance, Angwin et al (2016) note that Northpointe, the provider of COMPAS, "does not publicly disclose the calculations used to arrive at defendants' risk scores".
[37] Precisely how to satisfy this requirement for the workings of an ML tool is complex. It would not suffice to simply release the code that controls the algorithm's development, for there is an important difference between being 'open', by releasing as much information as possible, and being 'transparent', by facilitating genuine public understanding (O'Neal 2006).
[38] See Gardner (1998); Lippert-Rasmussen (2014); Segall (2012); and Shelby (2002) for critical discussion.
[39] See Lippert-Rasmussen (2014) and Eidelson (2015) for critical discussion, and Glasgow (2015) for a response to Lippert-Rasmussen.
[40] Some might make the stronger claim that, even if no-one knew about or suspected the presence of the proxy, perfect proxy profiling would have a disrespectful social meaning in this context. We remain neutral on this claim, but note that, given what is already known about machine learning, the presence of a perfect proxy is likely to be suspected by at least some members of any society that employed perfect proxy profiling.
[41] Adam Omar Hosein [personal communication] rightly notes that saying 'I would never be friends with someone from Brixton' plausibly expresses disrespect for Black people even though the population of Brixton is far from 100% Black.
[42] For discussion, see Hellman (2020, 848-53).
[43] Corbett-Davies et al. (2017, 805): "[E]xplicitly including race as an input feature raises legal and policy complications, and as such it is common to simply exclude features with differential predictive

power." Huq (2019, 1057): "the use of racially bifurcated thresholds would raise constitutional concerns akin to those engendered by affirmative action programs".

[44] Although she acknowledges (2020, 848-53) that it will be unlawful in cases where the threshold for high risk is set differently for different racial groups. See also Huq (2019, 1098).