Three lessons for and from algorithmic discrimination[i]

Abstract. *Algorithmic discrimination has rapidly become a topic of intense public and academic interest. This article explores three issues raised by algorithmic discrimination: 1) the distinction between direct and indirect discrimination, 2) the notion of disadvantageous treatment, and 3) the moral badness of discriminatory automated decision-making. It argues that some conventional distinctions between direct and indirect discrimination appear not to apply to algorithmic discrimination, that algorithmic discrimination may often be discrimination between groups, as opposed to against groups, and that it is not necessarily the case that morally bad algorithmic discrimination gives us reason to not use automated decision-making. For each of the three issues, the article explores implications for algorithmic discrimination, suggests some alternative answers, and clarifies how we may want to think of discrimination more broadly in light of lessons drawn from the context of algorithmic discrimination.*

Frej Klem Thomsen

Senior Consultant, Danish National Centre for Ethics

fkt@dketik.dk

1. Intro

Algorithmic discrimination has become a topic of rapidly increasing academic interest in recent years. This interest follows in the wake of a number of high-profile cases of intuitively problematic automated decision-making, such as the prominent COMPAS/ProPublica-debate, as well as the much-discussed impossibility theorems, which demonstrate that not all forms of apparently appealing equality are simultaneously realisable in realistic scenarios.[1]

[1] On the former, see Angwin et al. 2016, Flores et al. 2016, Dressel & Farid, 2018, Dieterich et al. 2016, Larson et al. 2016. On the latter, see Kleinberg et al., 2016, Chouldechova, 2017; for discussion, see Berk et al. 2018, Corbett-Davies & Goel 2018, Hedden, 2021, Heidari et al. 2019, Mitchell et al. 2021.

Academic analyses of discrimination and particularly of the ethics of discrimination have become increasingly sophisticated over the past two decades.[2] However, it is only very recently that philosophers and legal scholars have begun to contribute to the debate on algorithmic discrimination with insights grounded in those analyses. (See for example Davies & Douglas 2022, Lippert-Rasmussen 2022, Thomsen 2022, Hedden, 2021, Zimmermann & Stronach 2021; Hellman 2020, Castro 2019; Chiao 2019, Binns 2018; for an overview, see Fazelpour & Danks 2021)

One of the most important tasks in these contributions is the application of developed normative theory explaining the moral badness of discrimination to reviewing the fairness criteria developed in data science. There are, however, important further issues raised by algorithmic discrimination, many of which have yet to receive serious scrutiny. In this paper, I mean to discuss three such issues, each of which complicates moral analysis of algorithmic discrimination. On each of the three issues I believe there are lessons we can learn - in some cases from algorithmic discrimination for our general moral theory of discrimination, and in others from moral theory for how we evaluate algorithmic discrimination.

The first issue concerns the difference between direct and indirect discrimination. Briefly, it is often held that a) direct discrimination is intentional, whereas indirect discrimination is unintentional, and b) direct discrimination is, at least generally speaking, worse than indirect discrimination. I argue that neither of these differences appear to apply in the context of algorithmic discrimination, and that analysis of algorithmic discrimination might thereby help us to better understand how to distinguish between direct and indirect discrimination.

The second issue concerns the notion of disadvantage. It is commonly held that discrimination against a group involves treating them in a way that disadvantages members of the group. I argue that it is often unclear, in cases we might at first glance want to label algorithmic discrimination, that algorithmic analysis and its output in itself imposes a disadvantage on a group, which suggests that algorithmic discrimination is not discrimination *against* the group, and that any discrimination

[2] Prominent contributions include Collins & Khaitan 2018, Eidelson 2015, Hellman 2008, Hellman & Moreau 2013, Khaitan 2015, Lippert-Rasmussen 2013, 2018b, and 2020, Moreau 2020.

against the group might more accurately be said to occur in the context of the way the algorithm or its outcome is used. Here it seems to me that we might at once learn something from moral theory about algorithmic discrimination, and something from algorithmic discrimination about the importance of indirect discrimination.

The third issue concerns when discrimination gives us reason to not use algorithmic decision-making (ADM). It is sometimes held that an ADM is morally bad if it wrongfully discriminates, and sometimes that an ADM is morally bad only if it is more wrongfully discriminatory than a human alternative. I argue that both of these views have very counter-intuitive implications, and suggest that discrimination-based reasons against ADM rest on a more complex counterfactual comparison. Here it seems to me that we might learn something about algorithmic discrimination from moral theory.

The article proceeds in roughly the order above. In section two below I briefly sketch ADM and discrimination. In sections three to five I discuss each of the three issues in turn, drawing the lessons that it seems to me we might learn for and from algorithmic discrimination. Section six summarizes and concludes with some perspectives on implications for the broader debate.

2. Some notes on ADM and discrimination

This article is about ADM and discrimination. As such, I shall say a little about what I mean by those terms. I shall say no more than a little, in part because there are a great many other things to occupy us, and in part because there are already a great many such introductions to the topics in the literature. (e.g. Hacker 2018, Huq 2019, Kleinberg et al 2019, Chouldecova & Roth 2018, Hellman 2020, Barocas et al 2021) The objective here is only to ensure that the conception of these two key terms employed in this article is clear to the reader.

By ADM, I mean a software algorithm designed to deliver an output that can either form the basis of or support a decision. Such algorithms are instructions for performing a series of logical operations on input data. Their purpose is, broadly speaking, to perform a statistical estimation, by relying on correlations between the input data and a target value, such that we get the best guess possible of what the target value might be on the basis of the data we have. That statistical estimation is then the basis of a recommendation or decision, such as “loan applicant likely to default – deny loan”.

ADM are today often developed by machine learning, where a learning algorithm trains the ADM through fitting its mathematical function to a large set of historical data. Learning from such data – in a strictly technical sense – the mathematical function at the heart of an ADM is optimized for some specific definition of value, such as a measure of successful prediction.

As is often noted, the use of ADM is becoming rapidly more widespread across numerous sectors including health care (diagnosis), education (student and teacher assessment), marketing (profiling for advertisement), finance (credit assessment), social services (aid eligibility and child vulnerability), and even criminal justice (risk of recidivism). (See e.g. O'Neil 2016, Eubanks 2018) Given the accelerating use of ADM and their potential impact on important elements of modern life, it is evidently important to consider the ethics of using them. Perhaps the most prominent such ethical issue has been the risk of algorithmic discrimination.

What then of discrimination? A simple, provisional definition of direct discrimination against a group of persons loosely based on the work of Kasper Lippert-Rasmussen is the following:

> **Direct discrimination.** An agent *A* discriminates against (persons from) group *G* in a particular dimension D *iff* i) *A* treats *G* differently in some dimension *D* than she treats, or would have treated others, ii) the differential treatment is disadvantageous in *D* to *G*, and iii) the difference in treatment is suitably caused by group membership of *G*. (cf. Lippert-Rasmussen 2006; Lippert-Rasmussen 2013, p.14-22).

A few remarks on even this simple definition may be in order. First, I mean for us to interpret "agent" very liberally. Often, the discriminator will be an actual agent, but the definition also applies to laws, policies or practices, which can hardly be said to be conventional agents. More to the point, an ADM is not an agent in the standard sense, but it seems perfectly sensible to understand algorithmic (direct) discrimination along the lines above.

Second, the sense of discrimination at stake here is generic, in that it does not restrict discrimination to a particular set of "protected" or "socially salient" groups, such as those defined by gender, ethnicity, sexuality, religion or disability. There can, I believe, be reasonable disagreement about

whether discrimination is best understood generically or as pertaining to particular groups.[3] However, the choice here is strictly pragmatic. For present purposes, it will be simpler to discuss the issues while using a generic definition, and doing so does not preclude the possibility that discrimination should ultimately be understood as restricted to particular groups.

Third, the sense of discrimination is descriptive or non-moralised, in that it does not require an act to be in any way morally bad for it to constitute discrimination (which is compatible, of course, with discrimination sometimes or even always being in fact morally bad).

Fourth, the definition pertains to discrimination *against* a group (G), as opposed to *in favour* of the group or *between* the group and others, because it disadvantages the group. We shall return to and elaborate upon the difference between discrimination against and discrimination between in section four below.

Fifth, that the definition concerns discrimination in a particular dimension (D) means that it pertains to some particular good, such as e.g. employment, voting rights, or health care. Thus, an employer may discriminate against a group in the dimension of employment, in the sense of putting them at a disadvantage with respect to employment at her workplace, without us having said anything about how this affects other goods or a whole-life assessment of the discriminatee.[4]

Finally, the definition employs the unsatisfactorily vague qualifier that the difference in treatment be *suitably* caused by group-membership. Some condition of causal relation is obviously necessary in order to rule out labelling random distributions of disadvantage as discrimination, e.g. a case where a lottery happens to produce only winners from one group and only losers from another. A *suitable* causal relation is necessary in order to rule out labelling as direct discrimination counter-

---

[3] Cf. Lippert-Rasmussen 2013, Thomsen 2013.

[4] We can imagine cases where discriminating against a person (in a particular dimension) leaves her all things considered better off, such as Kasper Lippert-Rasmussen's example of an employer in Nazi Germany who refuses to hire Jews, and thereby inadvertently induces some Jews to migrate to the US prior to the Holocaust. Plausibly, it is all things considered better for these persons to be refused employment, immigrate and survive, than to be employed only to be killed, but we would still want to characterize the employer's treatment of them as discrimination against Jews (in the dimension of employment). (Lippert-Rasmussen 2013)

intuitive cases with strange or inappropriately downstream causal chains, such as where the decision to hold the lottery from above was itself in some perhaps very small way caused by gender. Ideally, we would sharpen this condition, but for the purposes of this article we shall restrict our attention to one particular way of doing so. Immediately below, we will review (and reject) the suggestion that differential treatment is suitably caused by group membership when and if the discriminator *intends* to discriminate on the basis of that group membership. Apart from this review, I shall assume that generally speaking an ADM that employs group membership as an effective feature in its function satisfies the condition, while an ADM that employs only features statistically correlated with group membership does not. As long as the reader is willing to grant these assumptions, we can set aside the broader and more complex question of how to make the condition precise.

Although much more could be said about both ADM and discrimination, these rudimentary remarks must suffice. Let us turn to the first of the issues, which concerns the distinction between direct and indirect discrimination, and the lessons we can draw from algorithmic discrimination on that issue.

3. Some potential differences between direct and indirect algorithmic discrimination

The most common sense of discrimination is direct discrimination. The definition above is intended to provisionally capture this sense. However, in both theory and practice we often distinguish a different sense of discrimination. Compare for illustration the following two cases:

> **Misogynist.** An employer refuses to hire women.
>
> **Shortist.** An employer refuses to hire anyone below average height.

Misogynist is, of course, a standard case of direct discrimination against women. The employer's treatment in *Shortist* is also disadvantageous to women, in the sense that far fewer women than men will have the opportunity for employment, for the simple reason that far more women than men are shorter than average (when we average height for the total population). *Shortist*, however, is not a conventional case of direct discrimination. It is rather an example of what we tend to label

indirect discrimination against women.[5] An obvious and important question is what exactly the difference between direct and indirect discrimination is meant to be? Unless we have an answer for that question, we will be unable to provide a clear definition of either direct or indirect discrimination.

Two answers that might seem appealing are i) that direct discrimination is *intentional* while indirect discrimination is unintentional, and/or ii) that direct discrimination is, perhaps in virtue of the first distinction but all else equal, morally *worse* than indirect discrimination. These answers need not be exhaustive – there could be further differences – but they are often supposed to be at least some of the most important differences between direct and indirect discrimination. However, analysis of algorithmic discrimination suggests that neither of the two constitutes a genuine distinction between direct and indirect discrimination, and that we must therefore look elsewhere in order to distinguish the two. Let us see how and why.

The first idea is that direct and indirect discrimination differ with respect to the agent's intentions. As noted above, this is one way of sharpening the condition that the differential treatment in direct discrimination be suitably caused by group membership. This distinction is prominently represented in anti-discrimination law, particularly in the US legal system, but is sometimes thought to apply not just in positive law but to the concept of discrimination. As Andrew Altman writes: "... an act that imposes a disproportionate disadvantage on the members of a certain group can count as discriminatory, even though the agent has no intention to disadvantage the members of the group and no other objectionable mental state, such as indifference or bias, motivating the act. This form of discriminatory conduct is called "indirect discrimination" or, in the language of American doctrine,

[5] Does *Shortist* satisfy the definition of (direct) discrimination I introduced in the previous section? After all, women are disadvantaged, and their disadvantage is caused by group membership? No, because although the *disadvantage* is caused by group membership, the *differential treatment* is not – the difference in treatment relies only upon the decision criterion 'height'.

"disparate impact" discrimination."[6] (Altman 2020) On this understanding, an act is thus only indirect discrimination if it involves no relevant, objectionable mental state.

On the other hand, an act that does involve an agent with an objectionable mental state directed at the discriminatee is *direct* discrimination, irrespective of whether the agent distinguishes on the basis of that group membership in her treatment of others. Altman cites as an example in support of this idea the use of literacy tests to disenfranchise Black voters in Jim Crow USA. Such tests did not, of course, directly distinguish on the basis of race: "Notwithstanding the absence of an explicit reference to race in the literacy tests themselves, their use was a case of direct discrimination. The reason is that the persons who formulated, voted for, and implemented the tests acted on maxims that *did* make explicit reference to race." (Altman 2020)

So, on the **broad mental state distinction**:

> Any discrimination that is based on an agent's objectionable mental state about the discriminatee is direct discrimination; any discrimination that is not based in this way on an objectionable mental state is indirect discrimination.[7]

Whatever its immediate appeal, the broad mental state distinction seems to me upon reflection to be mistaken. As Benjamin Eidelson has argued, we should distinguish between any discrimination involved in the decision to pursue a course of action (or to create or sustain a policy, practice or ADM), and any discrimination involved in that action itself. (Eidelson 2015) This is perhaps best brought out by cases that involve failed attempts to disadvantage a particular group. Consider:

---

[6] As Altman notes here, the distinction is arguably best conceived as including a somewhat broader set of objectionable mental states. Andrew Altman further observes that: "A disadvantage might [...] be imposed as a result of a general indifference toward the interests and rights of the members of a certain group. [...] Such instances of discrimination [...] should be counted as forms of direct discrimination, because the disadvantageous treatment derives from an objectionable mental state of the agent." (Altman 2020) I shall restrict the focus of the discussion to intentions for the purely pragmatic reason that this will be simpler, but the points made generalize, I believe, to other relevant mental states.

[7] Note that on some views, to wit disrespect-based accounts of discrimination, the mental state distinction might entail that direct discrimination is necessarily morally bad.

> **Inept misogynist.** An employer aims to avoid hiring female employees. To achieve this, he implements a policy of preferring applicants from particular schools and with particular stated hobbies, both of which he expects to correlate with gender. However, his understanding of these correlations is very outdated. As a result, the policy does not actually disadvantage women.

On the broad mental state distinction, we would be forced to say either that this case involves no discrimination, because there is no disadvantage, or that the policy directly discriminates against women, because of the employer's intentions. Neither conclusion sounds right. It seems to me more accurate to say that *Inept Misogynist* involves two distinct occasions of (potential) discrimination: direct discrimination, in the choice of policy, and a policy that was intended to be, but was not in fact, indirectly discriminatory. Presumably, similar cases where the misogynist is less inept will thus contain both direct discrimination in the choice of policy and indirect discrimination in the shape of the policy itself, which suggests that the intentional/unintentional distinction (or some broader form of mental state distinction) cuts across direct and indirect discrimination.

Proponents of a mental state distinction could accept this observation and revise the condition accordingly. If it is not sufficient for discrimination to be based on intention, in the downstream causal sense at stake in Inept misogynist, then perhaps the intention has to be to employ group membership as a distinguishing criteria. Hence, on the **narrow mental state distinction**:

> Any discrimination that is based on an agent's objectionable mental state about the discriminatee in the particular sense that this objectionable mental state causes the agent to employ membership of the group that defines discriminatees as a distinguishing criterion for differential treatment is direct discrimination; any discrimination that is not based in this way on an objectionable mental state is indirect discrimination.

Is this a more plausible distinction? Here, consideration of algorithmic discrimination can help us to answer that question. To illustrate, consider the following scenario: suppose that in some particular setting there is a measured correlation between gender and job performance. We can assume that this correlation is spurious, e.g. because it reflects discrimination of women in the measurement of job performance, or that it is true, e.g. because women are socialised in a way that makes them on

average worse at performing this type of job. I do not think this makes any difference to the argument at stake, so I invite the reader to adopt their preferred assumption. Suppose also that there are additional features that correlate with job performance *only* because these features correlate with gender. That is, height might correlate with job performance, but not because height independently predicts job performance, only because gender does and height correlates with gender. Call these features gender-proxied predictors. Now, compare four cases:

> **ADM 1.** An employer trains an ADM to screen applicants. He intends to avoid hiring women, and for that reason includes "gender" as a feature in the model. As a result, female applicants are less likely to be hired.
>
> **ADM 2.** As *ADM 1*, except the employer includes gender-proxied predictors as features, but does not include "gender".
>
> **ADM 3.** As *ADM 1*, except the employer has no discriminatory intentions. Gender is included as a feature without the employer being in any way aware that this will lead to discrimination.[8]
>
> **ADM 4.** As *ADM 2*, except the employer has no discriminatory intentions. Gender-proxied predictors are included without the employer being in any way aware that this will lead to discrimination.

It seems to me clear that in all of the cases, the ADM discriminates in some form.[9] However, the question is: in which of these cases is the ADM respectively directly and indirectly discriminatory? On the broad mental state distinction, ADM 1 and ADM 2 would both be direct discrimination, and

---

[8] This scenario is unlikely, of course, in that a developer will normally only include features she expects might correlate with the target (job performance), and is unlikely to be unaware that including gender as a feature can lead to discrimination. Furthermore, given the social importance of gender equality, we would expect a developer to pay particularly close attention to the use of gender as a feature. Unlikely as the scenario is, however, it is not impossible.

[9] Here, some may feel that it makes a difference whether the correlation is spurious or not. If so, I again invite the reader to adopt the assumption that will allow her to assume that all four cases involve discrimination against women (in some form).

ADM 3 and ADM 4 would be indirect discrimination. On the narrow mental state distinction, ADM 1 would be direct discrimination, and the remainder all indirect discrimination. Is this right?

Unquestionably, the employer intends to discriminate in *ADM 1* and *ADM 2*, and his *choice* of features is a form of direct discrimination, but what we are considering is what to say of the ADM *itself*. On that matter, it would seem to me very strange to say that the nature of the ADMs' discrimination is in any way affected by the employer's discriminatory intentions (or lack thereof) when developing it. The question of what type of discrimination the ADM performs seems to me entirely settled by the fact of which features the ADM employs: *ADM 2* and *ADM 4* indirectly discriminate, because of their use of gender-proxied predictors, while *ADM 1* and *ADM 3* directly discriminate, because of their use of gender itself as a feature. If that is correct, both the broad and the narrow versions of the mental distinction fail.[10]

What of the second conventional difference between direct and indirect discrimination? It is often held, even if not always explained, that the former is worse than the latter. We arguably find this notion built into discrimination law, where indirect discrimination is typically subject to a proportionality condition, such that cases that would constitute illegal indirect discrimination had they not met the condition, are legally allowed (i.e. legally considered to not be discrimination) if they pursue a legitimate aim, and the effect on that legitimate aim is proportionate to the disadvantage created. For example, in EU Council Directive 2000/43/EC, article 2, 2. (b): "indirect discrimination shall be taken to occur [...] unless [the apparently neutral, effectively disadvantageous] provision, criterion or practice is objectively justified by a *legitimate* aim and the means of achieving that aim are *appropriate* and necessary." The proportionality condition might, for example, make it permissible for an employer to hire only warehouse staff with a forklift-operating certificate, even if very few or no female potential workers hold such a certificate.

---

[10] Could one object that since gender causes the differences in values for the proxy-predictors, ADM 2 and ADM 4 are in fact cases of *direct* discrimination on the definition I have offered? Not on the intended reading of condition iii) ("the difference in treatment is *suitably* caused by group membership…"). Although I have set aside as going (far) beyond the scope of this article the task of sharpening that condition, it seems clear that the line would have to be drawn in a way that precludes such downstream causation from being a suitable cause.

The inclusion of a proportionality condition raises the bar for prohibited indirect discrimination. There will be some cases of discrimination that would be prohibited had they been direct discrimination, but which are permitted because they are indirect discrimination, because the latter can meet the proportionality condition while the former cannot.[11] Why this difference?

The most obvious way to explain the distinction is that direct discrimination is all else equal, or at the very least generally speaking, worse than indirect discrimination.[12] It is worth emphasizing here that one cannot infer from discrimination law when discrimination is morally wrong, nor can we reason the other way and say, for example, that discrimination that is morally wrong is for that reason unlawful, nor even that it ought (all-things-considered) to be. Positive law and morality are different fields, not unrelated, but the relations between the two are complex and indirect. So my claim is not that the legal proportionality condition shows that there is a moral difference between direct and indirect discrimination, only that the idea that there is a moral distinction can explain why legislators and judges have been motivated to draw a particular legal distinction. That is, the legal distinction *suggests* that some *believe* that direct discrimination is all else equal worse than indirect discrimination.[13] Call this **the moral distinction**:

> Direct discrimination is all else equal, or generally speaking, morally worse than indirect discrimination.

---

[11] Could we say instead that the proportionality condition means that indirect discrimination will if anything be worse than direct discrimination? After all, if two cases of direct and indirect discrimination are equally bad in other respects, but the indirect discrimination is also disproportionate, does that not make it worse? This strikes me as a mistake. The proportionality condition is a way for certain cases of discrimination to become permissible. The presupposition is that direct discrimination is so inherently morally bad that not even the proportional pursuit of a legitimate aim can justify it, whereas indirect discrimination, *because* it is less wrong, can sometimes be justified on these grounds. Or, which is perhaps putting the same point differently, that indirect discrimination is potentially proportionate, whereas direct discrimination is incapable of being proportionate.

[12] Strictly speaking, the claim must be that direct discrimination is all else equal worse, for there to be a categorical difference that could explain why only indirect discrimination can be justified by its being proportional. It seems charitable, however, to assume that the legal prohibition might track only a tendency.

[13] My experience is that, at least in Western Europe and the Anglophone world, this view is quite widespread both with legal scholars and more broadly.

Assuming that we have stronger reason to prohibit an act, the worse the act is, this would explain why it might be possible to justify indirect discrimination but not direct discrimination.[14] Should we accept the moral distinction?

There are, I believe, general reasons to be skeptical of the claim (cf. Lippert-Rasmussen 2014, Thomsen 2015), but it is worth pursuing it in the particular context of algorithmic discrimination, since as with the mental state distinction above there are characteristics of this context that help to illuminate the issue.

A first suggestion might be that direct discrimination is all else equal/generally (I shall mostly omit this qualifier) morally worse because it involves an objectionable mental state, and acting on such mental states makes discrimination worse. This suggestion faces two immediate difficulties. The first difficulty is that it is doubtful that the presence of objectionable mental states makes an act of discrimination morally worse. Plausibly, such mental states affect how we ought to assess the moral character of the agent, but not the moral status of the act. (Lippert-Rasmussen 2013, Lippert-Rasmussen 2018a, Thomsen 2023; cf. Scanlon 2008) This position seems even more appealing in the context of evaluating algorithmic discrimination. An ADM, we commonly assume, is patently incapable of mental states of any kind, such that if we want to consider whether it discriminates in a morally problematic way, it is not readily apparent how mental states could form part of the explanation.[15]

The second difficulty is that, as we have seen above, it is not clear that the difference between direct and indirect algorithmic discrimination is best understood as resting on the presence of

---

[14] A related, alternative response could be that there are generally stronger reasons in favour of indirect discrimination than there are in favour of direct discrimination. But it is far from clear that the claim is true – there may be many situations in which there are fairly strong reasons to directly discriminate (Wertheimer 1983), and even more situations in which reasons to indirectly discriminate are quite weak.

[15] Perhaps pan-psychism is true (I do not think it is), in which case ADM has mental states because everything does. Or perhaps, as a reviewer suggested, there are non-conscious mental states, such as belief-like states, that at least some ADM can possess (I do not think there are). Perhaps there will one day be AI that has mental states (I think there likely will be). I set aside these possibilities here, however, and proceed on the assumption that ADM does not have mental states of any relevant kind.

objectionable mental states. This presents the proponent of the moral distinction with a dilemma: on the first horn, she insists on applying the mental state distinction. This has the awkward result that she cannot say that *ADM 3* is worse than *ADM 4*, since neither involves objectionable mental states, and must therefore accept that *ADM 3* should equally be subject to the proportionality condition. This will likely strike many attracted to the moral distinction as an unappealing prospect, since it means abandoning the notion that algorithmic discrimination that employs certain sensitive features is for that reason particularly objectionable. (cf. Grgic-Hlaca et al. 2018) Alternatively, the proponent concedes that the mental state distinction is inadequate. On that horn, the *ADM 1-4* cases amply illustrate that direct and indirect discrimination can equally involve objectionable mental states, and that the effect of objectionable mental states can therefore not support the moral distinction.

An alternative, and arguably more promising suggestion, is that direct discrimination is morally worse because it is a stronger form of discrimination, in the particular sense that the property that defines the group plays a greater part in causing the disadvantage. Direct discrimination is stronger in this sense, some might say, because it will necessarily affect all members of the targeted group, while indirect discrimination will typically only affect some. If discrimination is morally worse the stronger the discrimination, then direct discrimination will be at least as bad as and typically worse than indirect discrimination.

Let us accept for the purposes of argument the idea that discrimination is worse the stronger it is. Having done so, algorithmic discrimination readily illustrates that the other premise is false: direct discrimination is not necessarily as strong as and typically stronger than indirect discrimination, and as such it need not be as bad as or worse than indirect discrimination. Consider:

> **Simple classifiers.** An ADM, SC1, employs two features, X_1 and X_2, to predict outcome Y_1. X_1 is a numerical feature, scaled to vary between 0 and 1. X_2 is a categorical feature, "gender", assigned values 0 for men and 1 for women. Another ADM, SC2, employs features, X_1 and X_3, to predict outcome Y_2. Like X_1, X_3 is a numerical feature, scaled to vary between 0 and 1. Unlike X_1, however, X_3 is a gender-correlated predictor, such that women tend to have values closer to 1 and men to have values closer to 0. The models also employ weights for each of the

features, β_1, β_2, and β_3 (as well as an intercept, β_0, which we shall ignore), such that the function for each feature is the product of the weight and the value that the feature assumes. The weights defined are β_1: 100 β_2: 2, and β_3: 5000.

SC1 and SC2 are very strange models that we would not expect to encounter "in the wild"; they serve only to make a point. That point, however, will also apply to real cases. The point is this: the effect discrimination on the basis of any feature X_n has on any particular group G depends not merely on the proportion of G that is affected, but also on the size of the effect X_n has on the outcome. This is easy to see when we consider how gender affects the outcomes of SC1 and SC2 respectively. For SC1, the effect of gender is dwarfed by the other feature whose weight is 50 times greater. Setting aside cases with very unusual distributions of values in X_1, gender may rarely or never make any difference to the outcome. For SC2, the situation is reversed – the weight of X_3 is 50 times as great as X_1. The effect of gender may be enormous – so big as to all but determine the outcome irrespective of what values X_1 assumes.[16]

This seems to me an often overlooked point, particularly since it applies to human decisions too. Discrimination can have very little effect on a decision, and be discrimination all the same. But if we ought indeed to be more concerned with discrimination the stronger it is, then we will need to consider the magnitude of its effect. Once we do so, it is apparent that there is no reason to think that direct discrimination will necessarily or even in general be stronger than indirect discrimination.

The first issue of this article has been the difference between direct and indirect discrimination. Insights from analysis of algorithmic discrimination, I have argued, can help challenge some of the conventional ways of understanding the difference between the two forms of discrimination. I have

[16] There are cases, of course, where direct discrimination necessarily exerts a greater influence on the outcome than indirect discrimination. Two algorithms predicting the same outcome and optimized for accuracy will not ascribe greater weight to a suitably scaled, gender-correlated feature than to gender since the former is a proxy predictor of the latter. This is compatible with the claim at stake, however, that we cannot know the size of the effect a feature exerts on the outcome unless we know the weight it is assigned and how that weight compares to the weights assigned to other features of the algorithm, and that therefore we cannot say that direct discrimination is necessarily worse than indirect discrimination because more strongly discriminatory. I am grateful to an anonymous reviewer for pressing me on this point.

argued above that there are conceptual difficulties with the mental state distinction as a means of distinguishing direct and indirect discrimination. Algorithmic discrimination helps illustrate this by providing cases where the presence or absence of the relevant mental states appear not to affect what we intuitively want to label direct and indirect algorithmic discrimination. A similar point applies to the moral distinction, where at least two of the most obvious ways of supporting the claim that there is a moral difference between the two turn out to be misguided, in a way that is helpfully illustrated by cases of algorithmic discrimination. There may be other, contingent moral differences, of course – direct discrimination may be more likely to cause expressive harm, for example – but these are better conceived as differences between discrimination (direct or indirect) that is contingently bad (e.g. causes expressive harm) and discrimination that is not.

It remains an open question exactly how to draw the conceptual distinction on different grounds, even if the analyses of *ADM 1-4* has at least hinted at an answer. (For further discussion, see Thomsen 2015) And it is of course possible that there will turn out to be moral differences between the two on our best understanding of the conceptual difference. Yet it seems to me that one of the lessons we can learn from algorithmic discrimination is that direct and indirect discrimination are neither conceptually nor morally as distinct as is sometimes assumed.

4. The disadvantage of algorithmic discrimination

We have noted above that conventional cases of discrimination against a group concern differential treatment that disadvantages the group. This distinguishes discrimination against a group from discrimination in favour of a group and discrimination between groups, which respectively advantage the group and neither advantages nor disadvantages the group. Consider for illustration:

> **Sexist gym.** A gym provides separate and better shower and changing facilities for men than for women.

> **Segregated gym.** A gym provides separate but equally good shower and changing facilities for men and women.[17]

Discrimination against and discrimination in favour are symmetrical, in that discrimination against a group can as accurately be called discrimination in favour of the inverse ("all but the group"), and vice versa. Thus, conceptually it makes no difference whether we describe *Sexist gym* as a case of discrimination in favour of men (who are provided better facilities than women), or discrimination against women (who are provided worse facilities than men), even if there are sometimes pragmatic reasons for preferring one of the two descriptions.

Discrimination in favour of and against groups differ, however, from discrimination *between* groups in that the latter does not involve imposing advantages or disadvantages. One question we might ask of algorithmic discrimination is therefore this: does it typically involve discrimination between groups or discrimination against one group in favour of another?

The answer, I want to argue, is less obvious than it might appear. Certainly, there are conceivable cases where algorithmic discrimination might be said to impose disadvantage. Consider:

> **Decision bias.** A classification model is employed to distribute an important good. The model calculates a metric of deservingness, and then employs a human-defined decision threshold: below the threshold, claims for the good are denied,

[17] Assume for simplicity that the cases occur in a society with only these two genders. Segregated gym could arguably involve discrimination against other gender groups, but for present purposes I want to set such complications aside. Note also that we assume that the separate facilities are *actually* equally good. A central problem with historical examples of "separate but equal"-policies was that they lived up to only the first half of that label. Furthermore, discrimination between in a particular dimension, such as changing facilities, can be bad for a group in a different dimension, e.g. because the discrimination stigmatizes the group. To take a famous example, it is not clear that being a passenger in the front of a bus is intrinsically better or more desirable than being a passenger in the back of a bus, but a policy that requires a minority to sit at the back of the bus might harm that minority nonetheless by reinforcing false stereotypes, deepening social divisions, and undermining self-respect. However, while it is important to recognize these points when morally evaluating discrimination, they do not affect the conceptual point at stake here.

> while at or above the threshold they are approved. The decision threshold varies depending on the ethnicity of the claimant.

Here, the model differentially treats groups on the basis of ethnicity, and the differential treatment is disadvantageous because it makes it more difficult for some to access an important social good than it is for others, or putting the point differently, because it means that equally deserving individuals do not have equal opportunity to access the good.

As it happens, ADM rarely discriminates this way. First, ADM will typically predict a factual and easily measurable outcome, simply because these are the outcomes for which developers can obtain good training data.

Second, apart from special cases, such as where different decision-thresholds are used to reduce outcome disparities (cf. Lipton et al 2018), ADM will mostly employ a uniform decision threshold. The properties at the heart of our concerns about discrimination are more likely to appear (if at all) as standard features in the model. In *Decision Bias* this would mean basing the calculation of deservingness in part on ethnicity.

Why a uniform decision threshold? Because as we noted initially, ADM are ordinarily designed to provide a best guess at a target value. Their purpose is to provide information that can form the basis of or support a decision, and they are typically trained with a loss function that gives prime importance to optimizing the accuracy of the prediction. Now here's the rub: when ADM employs a uniform decision threshold, and is designed simply to accurately predict a property, it is far from obvious that the model can discriminate *against* as opposed to *between* groups. Consider:

> **Diagnosis.** A classification model is employed to diagnose patients. The model calculates the probability of a patient being positive for a dangerous disease, and then employs a human-defined decision threshold: below the threshold, the model returns a "healthy" diagnosis, while at or above the threshold it returns an "ill" diagnosis. Gender and ethnicity are features in the model, and as such are used to predict the probability of being ill.

Let us suppose that gender and ethnicity are features in the model for good reason. That is, employing these features allows more accurate predictions, because risk varies with gender and

ethnicity, or because other features interact with these features (e.g. advanced age increases probability of illness, but at different rates for men and women). Let us also assume that the diagnosis is used in something like a standard scenario, i.e. doctors offer treatment on the basis of the diagnosis, and it is better for those ill to receive treatment, and better for those healthy to not receive treatment. In this case, it would seem very strange to say that the ADM's differential treatment disadvantages any group.

Why does *Diagnosis* not involve disadvantage? The most obvious explanation is that the fact that the model discriminates between these groups leaves everyone better off than they would otherwise have been, and no group better off relative to another group.

A different explanation is possible, however. Suppose *Diagnoses* occurs in a racist and sexist dystopia, where authorities respond to diagnoses very differently: women and ethnic minorities are not offered treatment, but "euthanized" (i.e. murdered), if they are diagnosed as positive. Clearly this scenario involves gross disadvantage and horrible discrimination against these groups. The question is: does the ADM – the ADM itself – discriminate against them? On reflection, it is not clear that we can say so. After all, nothing about the ADM has changed between the two scenarios. It provides exactly the same, purely factual prediction. What has changed is the way this information is used. Thus, an alternative description of the first scenario is that the ADM does not impose disadvantage because the statistical prediction does not itself constitute an advantage or disadvantage – advantages and disadvantages are determined by the actions of the agent who decides what the result of that prediction will be. On that interpretation the ADM discriminates *between* groups in both and *against* groups in neither version of the scenario.

How would this interpretation apply to cases of indirect discrimination? Consider:

> **Recidivism tools.** R1 and R2 are ADM that predict risk of recidivism. There are differences in recidivism risk across ethnic groups. R1 employs ethnicity as a feature. R2 does not employ ethnicity, but does employ ethnicity-proxied predictors.

Intuitively and in line with the analysis in section three above, R1 directly discriminates while R2 indirectly discriminates. On the interpretation at stake, however, both discriminate *between* groups

as opposed to against. This is again rendered plausible when we consider that the information could be used in many ways. Authorities could, for example, use the information to provide special assistance and care to offenders judged at high risk of recidivism, in order to help them reduce the risk. Conversely, authorities can employ the information in what is sadly the more common way, to justify prolonged incarceration. But as in *Diagnosis*, this appears to be a difference in the use of the information, which should not affect how we describe the ADM's discrimination.

Does this mean that we cannot criticize the use of ADM to predict recidivism risk in such scenarios as discriminatory? Hardly. When there are differences in recidivism risk across ethnic groups, then the use of predicted recidivism risk to impose disadvantages, such as prolonged incarceration, will be *indirectly* discriminatory against statistically over-represented groups, whether the information obtains from R1 or R2. (Cf. Lippert-Rasmussen 2021) Notably, however, this is also the case when the risk assessment is made by humans.

Indirect discrimination between groups in ADM may already be more likely than direct discrimination, because of more stringent norms and stricter legal prohibition against direct discrimination. The interpretation at stake suggests that indirect discrimination against groups in the *use* of ADM may also be far more common than direct discrimination against groups, either in the ADM itself or in its use. If our goal is to reduce or prevent morally bad discrimination against vulnerable groups, we may want to focus more on the way ADM is used, and the indirect discrimination this use can involve, and less on differential treatment in the ADM.

5. The baseline for bad algorithmic discrimination

The third and final question that I wish to pursue is this: when does the fact that using ADM will involve morally bad discrimination give us reason to not use ADM? Or, for short, when does morally bad discrimination make use of an ADM morally bad?[18]

[18] Bear in mind that use of an ADM being bad, in the sense at stake, means only that there is a discrimination-based reason against using the ADM. This leaves open e.g. whether we ought all-things considered to employ the ADM. It is possible, for example, that there could be scenarios where other benefits of ADM, such as higher general accuracy or lower costs, justify some amount of bad discrimination. Whether or not we accept this possibility of

Two initial clarifications may be worth making. First, it may be worth emphasizing that this section does *not* ask what makes discrimination morally bad, or (which is to ask the same question of the specific context) when discriminatory ADM is morally bad. While the question of what makes discriminatory ADM morally bad has, as previously noted, already received a fair amount of scholarly attention, the question at stake here is currently underexplored. [19] In the interest of making the point broadly applicable, I shall remain agnostic about the issue of what makes discriminatory ADM morally bad, and allow the reader to assume their preferred theory, be it fairness, harm, disrespect or something else.

Second, posing this question may seem odd in light of the conclusion of the preceding section. I have argued above that ADM will often discriminate between rather than against groups. Intuitively, discrimination between groups is ordinarily less morally bad, if bad at all, than discrimination against a group. [20] The problem with algorithmic discrimination may therefore be mainly indirect discrimination in the use of ADM, as opposed to direct or indirect discrimination in the ADM itself. If we accept or assume this conclusion, does it make sense to explore when we have reason not to use ADM because doing so involves morally bad discrimination?

---

discrimination-based reasons being outweighed by other concerns, the issue remains when we have a discrimination-based reason against using ADM in the first place.

[19] Some accounts of moral badness focus on the effects that ADM will have in a particular context. It may be worth noting that the claim that the moral badness of discriminatory ADM hinges on the effects the ADM has in a particular context is different from the claim I will defend below, that whether we have moral reason to avoid using ADM that is morally bad on grounds of discrimination hinges on what the relevant alternatives are. I owe thanks to two anonymous reviewers for pressing me to clarify this point.

[20] This might be because it typically is less disrespectful (if disrespectful at all) to discriminate between groups (cf. Alexander 1992, Eidelson 2015, Slavny & Parr 2015; for critical discussion see Lippert-Rasmussen 2013, 2018b, Arneson 2013; Thomsen 2023), because it is typically less demeaning (if demeaning at all) to discriminate between groups (cf. Hellman 2008, Shin 2009; for critical discussion see Lippert-Rasmussen 2013, Arneson, 2013, pp. 91-94; Eidelson, 2015, pp. 85-90; Ekins, 2012), or because it is typically less harmful (if harmful at all) to discriminate between groups (cf. Lippert-Rasmussen 2006, 2013, Arneson 2017). Although I find the harm-based account compelling, and none of its competitors persuasive, I take no stand here on which (if any) of these accounts are true, or even on whether we should credit the intuitive difference between discrimination against and discrimination between (see also remarks in footnote 18 above).

Any apparent oddness notwithstanding, the question is perfectly sensible. First, although I have suggested that they are much less common than they are widely thought to be, there may still be cases of morally bad algorithmic discrimination. Perhaps there are cases of algorithmic discrimination against groups, as suggested by *Decision bias*, or perhaps there are cases where algorithmic discrimination between groups is morally bad. Second, the indirect discrimination at stake when using an ADM may be morally bad. I have been careful to pose the question so as to include such morally bad indirect discrimination.

The question might also appear odd because the answer might seem obvious: use of an ADM is morally bad, with respect to discrimination, whenever using the ADM discriminates in a morally bad way. Indeed, it might seem intuitive that such discrimination-based reasons are strong enough to be ordinarily decisive. In the famous COMPAS-debate, ProPublica and many of their allies appeared to adopt this stance: COMPAS discriminates in a morally bad way because its errors tend to be skewed towards overestimating risk of recidivism when assessing Black offenders and underestimating risk of recidivism when assessing White offenders, and *because* COMPAS is discriminatory in this way, we should abolish its use. (Angwin et al. 2016; for discussion, see Flores et al. 2016, Dressel & Farid 2018, Dieterich et al. 2016, Larson et al. 2016)

Proponents of this view apply what we can call the **no bad discrimination** baseline:

> Using ADM is morally bad whenever the use of ADM involves morally bad discrimination.[21]

This clarification may seem to reinforce the oddness of the issue at stake. How could it not be the case that use of an ADM that involves morally bad discrimination is for that reason bad? In fact, a closer look at COMPAS-like ADM will help illustrate that the issue is a good deal more complicated than it first appears.

Let us grant the assumption that the difference in error rates does indeed mean that COMPAS discriminates in a morally bad way. A convincing argument to that effect is more difficult to make

[21] Recall that discrimination is not by the definition I have stipulated necessarily morally bad, and that there appear to be cases of morally benign discrimination, as illustrated by *Diagnosis*.

than is often assumed, as illustrated by the lessons drawn on the second issue above, but I want to focus on the more general claim. Is it true that this discrimination makes (or would make) use of COMPAS morally bad, perhaps to the extent that we ought all-things-considered not to use it?

As several research groups have famously demonstrated, it is in realistic scenarios (i.e. imperfect accuracy) impossible to achieve both equal distribution of errors (equal ratio of false positives to false negatives) and equal accuracy measures across groups with unequal base rates. (Kleinberg et al. 2016, Chouldechova 2017) That is, COMPAS-type scenarios force developers of ADM to trade off two intuitively appealing forms of equality. An important further point, less often noted, is that since this follows mathematically from the facts of the situation, the necessary trade-offs are not limited to ADM but apply equally to any human decision-maker facing the same situation. That is, we cannot avoid tradeoffs in these respects by replacing ADM with a human decision-maker. This illustrates an intuitive difficulty for the *no discrimination* baseline, which is brought out even more starkly if we consider scenarios where the alternative to discriminatory use of ADM is *even more* discriminatory human decision-making. Consider:

> **Judicial bias.** A jurisdiction employs R2 (from *Recidivism tools* above) to predict recidivism risk for offenders. The alternative to R2 is letting human judges predict recidivism. Many judges are subject to strong implicit racial biases, and some are consciously and avowedly, although discreetly, racist.

Clearly, it is an open question whether any actual jurisdiction is like *Judicial bias* – we may hope that most are better, and worry that many are not – but the point at issue does not depend on the extent to which real life jurisdictions match the case.[22] The point is just this: the decision-makers in *Judicial bias* will combine the mathematically necessary trade-offs that R2 suffers from with serious psychological biases that it does not. As such, the human alternative will, we can safely assume, be

[22] Sadly, the scenario is all too likely to be realistic. An abundance of studies have demonstrated that judges are as human as the rest of us, in that they are subject to the biases and prejudices also found in non-judges. (E.g. Rachlinski et al. 2009, Kang et al. 2012, Liu and Li 2019, Chiao 2021)

more discriminatory than the ADM.[23] Can we in such situations say, as the *no discrimination* baseline holds, that using the ADM is morally bad, in the sense that we have discrimination-based reason not to use it? It seems clear that the answer must be no; the *no discrimination* baseline fails.

Can we come up with a more suitable replacement? In light of the concerns noted above, we might be tempted to employ the **less badly discriminatory than humans** baseline:

> Using ADM is morally bad if ADM is more badly discriminatory than human decisions in the same scenario.

The *less badly discriminatory than humans* baseline also encounters problems. Problems emerge, for example, when we consider situations where we can choose between different types of ADM that vary with respect to discrimination, but are superior to human decision-makers. Consider:

> **Judicial and ADM bias.** A jurisdiction employs R2 to predict recidivism risk for offenders. One alternative to R2 is letting human judges predict recidivism. Many judges are subject to strong implicit racial biases, and some are consciously and avowedly, although discreetly, racist. Another alternative to R2 is competing ADM R3, which is somewhat more badly discriminatory than R2, but less badly discriminatory than human decision-makers.

Do we have discrimination-based reason to not use R3? Intuitively, the answer is clearly yes. R2 is less badly discriminatory, and it seems obvious that this gives us reason to not replace it with R3. The *less badly discriminatory than humans* baseline, however, entails the opposite: Since both are better than human decision-making, we have no discrimination-based reason to not employ R3. This cannot be right, and as a result, the *less badly discriminatory than humans* baseline also fails.

---

[23] Dressel & Farid 2018 have shown that COMPAS, specifically, appears to perform no better than and in some respects worse than untrained human predictors. Does this make the assumptions of *Judicial bias* implausible? Not necessarily. Other recidivism risk prediction tools perform better than humans, particularly in some scenarios (see Kleinberg et al. 2018; Lin et al. 2020). And even were that not the case, the theoretical thrust of the case does not depend on any actual risk recidivism tool performing better than humans, only upon the fact that there can be situations in which we face the choice presented in *Judicial bias*.

The problem alluded to above is particularly pressing in the context of developing ADM. In recent years there has been an intense research-focus on technical methods for training ADM that reduce different measures of discrimination. (Chouldecova & Roth 2018, Friedler et al. 2019) As such, the form and amount of discrimination in an ADM is increasingly a design choice for developers. It would be clearly unsatisfactory if developers were to aim no higher than the human alternative, when a less badly discriminatory ADM is trainable at little or no cost.

The particular way the *less badly discriminatory than humans* baseline fails might then appear to point in the direction of the ambitious **least badly discriminatory ADM baseline**:

> Using ADM is morally bad if it is more badly discriminatory than an ADM we could train.

The *least badly discriminatory ADM* avoids the problem illustrated in *Judicial and ADM bias*, because any existing ADM is an ADM we could train, and as such the baseline entails that there *is* a discrimination-based reason not to employ R3. It also means that we will always have discrimination-based reason to train the least badly discriminatory ADM possible. However, the problems that beset it are at this point likely obvious. In one respect, the *least badly discriminatory ADM* sets the bar too low. It entails that there is no discrimination-based reason against using an ADM even when the human alternative would be less badly discriminatory, so long as the ADM is the least badly discriminatory ADM we can train. This cannot be right. In another respect, it sets the bar too high. The *least badly discriminatory ADM* entails that there is discrimination-based reason against using an ADM, even when the only immediately available alternative is a decision method (human or ADM) that is more badly discriminatory, so long as it is possible to train a less badly discriminatory ADM. This also cannot be right. The fact that it is possible to train a less discriminatory ADM would appear to give us reason to do so, but so long as the question is what to employ in the meantime, then it would seem bizarre to say that we have discrimination-based reason not to use an ADM even though doing so would *decrease* bad discrimination.

Ultimately, the baseline must be context-sensitive, in the sense that it compares any bad discrimination of an ADM with relevant alternatives. So I suggest we should adopt the **decision alternatives** baseline:

> Using ADM is morally bad if it is more badly discriminatory than an available alternative.[24]

The notion of what constitutes an available alternative is notoriously theoretically difficult, but will in many cases be uncontroversial. Roughly, in the use stage, an alternative is available if by deciding to use it we can actually use it, and in the development stage, an alternative is available if by deciding to develop it we can actually develop it. Given a suitable precisification, the context-sensitivity of this baseline allows it to avoid the problems we have encountered above, and to give credible answers in both the development and use stages.

6. Concluding remarks

Algorithmic discrimination is a complex phenomenon. Even more so, perhaps, than has been widely appreciated. This complexity is both a challenge and an opportunity. Thinking through some of the complications we encounter in the context of algorithmic discrimination can help us to understand not only it but also illuminate discrimination more widely.

Over the course of this article, I have raised three issues about algorithmic discrimination, and drawn lessons both for and from algorithmic discrimination. I have argued that some of the alleged differences between direct and indirect discrimination seem not to apply in cases of algorithmic

---

[24] As will be apparent to legal scholars, *Decision alternatives* resembles the necessity condition often applied in discrimination law, as when US law holds that indirect discrimination (disparate impact) can occur even when the discriminator has good reason for the discriminatory practice (business necessity), if an alternative practice exists which would have had less discriminatory results, or when EU law similarly holds that a measure can constitute indirect discrimination regardless of justification, if another measure exists that would fulfill the same aim and be less detrimental to the group at stake. (On the former, see Barocas and Selbst 2016, p.701; on the latter, see Gerard & Xenidis 2020, p.73). While the use of counterfactual comparisons is clearly similar, the issue at stake is different in that i) *Decision alternatives* explains when morally bad discrimination gives us *pro tanto* reason to avoid using ADM, whereas discrimination law imposes a legal constraint, and ii) discrimination law focuses on whether an alternative exists that would be *less disadvantageous* for the group at stake, whereas the issue here is whether an alternative exists that is *less badly discriminatory*, i.e. less morally bad on grounds of discrimination (on some accounts of what makes discrimination morally bad, the two will correlate, but not on all). I am grateful to an anonymous reviewer for introducing and pressing me to clarify these points.

discrimination, and that their failure to do so should help us reconsider how we conceive of the distinction more broadly.

I have also argued that algorithmic discrimination would seem to often concern discrimination between groups rather than discrimination against groups, and that in such cases it may be more helpful to focus on the indirect discrimination of using an ADM than on discrimination in the ADM. The general lesson from this issue is that we may under-appreciate the frequency and importance of indirect discrimination.

Finally, I have argued that it is not as clear as is sometimes assumed when an ADM instantiating or leading to morally bad discrimination makes use of the ADM morally bad. This, it seems, depends entirely on what the available alternatives are, since in some cases these may involve even worse discrimination. This too is a lesson that potentially applies to discrimination in other contexts – policy or legislation, for example, that involves discrimination may in some cases be less badly discriminatory than the alternative. Thus, even limiting our attention to discrimination-based reasons, evaluating the moral badness of a discriminatory act, policy, practice or ADM may require at least one further step than is sometimes assumed.

Some of the lessons can perhaps be extended in other ways. Beyond algorithmic discrimination, a strong concern with ADM has been the lack of transparency in "black box" algorithms. Scholars have already noted that such concerns appear to sometimes overlook the extent to which human alternatives to ADM lack transparency. (Zerilli 2019, Chiao 2021) The analysis of when uninterpretable ADM is morally bad could, I believe, proceed along much the same lines as our analysis of the third issue has followed above. But the full pursuit of such an argument must, like many others, remain the task of another day. The field of data ethics is still young and full of promise.

References

Alexander, L. (1992). "What Makes Wrongful Discrimination Wrong? Biases, Preferences, Stereotypes and Proxies." *University of Pennsylvania Law Review* 141: 149-219.

Altman, A. (2015). "Discrimination." *Stanford Encyclopedia of Philosophy*. E. N. Zalta (Ed.).

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

Arneson, R. J. (2013). “Discrimination, Disparate Impact, and Theories of Justice.” In: D. Hellman & S. Moreau (Eds.), *Philosophical Foundations of Discrimination Law:* 87-111. Oxford: Oxford University Press.

Arneson, R. J. (2017). “Discrimination and Harm.” In K. Lippert-Rasmussen (Ed.), *The Routledge Handbook of the Ethics of Discrimination:* 151-163. London: Routledge.

Barocas, S. & Selbst, A.D. (2016). “Big Data’s Disparate Impact.” *California Law Review* 104 (3): 671-732.

Barocas, S., Hardt, M., & Narayanan, A. (2021). *Fairness and machine-learning*. Retrieved from https://fairmlbook.org/

Berk, R., H. Heidari, S. Jabbari, M. Kearns and A. Roth (2018). "Fairness in Criminal Justice Risk Assessments: The State of the Art." *Sociological Methods & Research*, Online first.

Binns, R. (2018). "Fairness in Machine Learning: Lessons from Political Philosophy." *Journal of Machine Learning Research* 81: 1-11.

Castro, C. (2020). “What's Wrong with Machine Bias.” *Ergo - An Open Access Journal of Philosophy* 6 (15): 405-426. doi.org/10.3998/ergo.12405314.0006.015.

Chiao, V. 2019. "Fairness, accountability and transparency: notes on algorithmic decision-making in criminal justice." *International Journal of Law in Context* 15 (2): 126-139.

Chiao, V. (2021). “Sentencing and the right to reasons.” In: Ryberg, J. & Roberts, J. (Eds.) *Sentencing and Artificial Intelligence*. Oxford: Oxford University Press.

Chouldechova, A. (2017). "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments." *Big Data* 5 (2).

Chouldechova, A. and A. Roth. 2018. "The Frontiers of Fairness in Machine Learning." *arXiv e-prints*.

Collins, H., & Khaitan, T. (Eds.). (2018). *Foundations of Indirect Discrimination Law*. Oxford: Hart Publishing.

Corbett-Davies, S., & Goel, S. (2018). “The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning.” *arXiv e-prints*. Retrieved from https://arxiv.org/pdf/1808.00023.pdf

Davies, Benjamin & Douglas, Thomas (2022), “Learning to Discriminate: The Perfect Proxy Problem in Artificially Intelligent Criminal Sentencing.” In: Ryberg, J. & Roberts, J. (Eds.) *Sentencing and Artificial Intelligence*. Oxford: Oxford University Press.

Dressel, J. and H. Farid. 2018. "The accuracy, fairness, and limits of predicting recidivism." *Science advances* 4 (1)

Eidelson, B. (2015). *Discrimination and Disrespect*. Oxford: Oxford University Press.

Ekins, R. (2012). “Equal Protection and Social Meaning.” *The American Journal of Jurisprudence* 57(1): 21-48.

Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police and Punish the Poor*. New York: St. Martin's Press.

Fazelpour, S. & Danks, D. (2021). “Algorithmic bias: Senses, sources, solutions.” *Philosophy Compass* 16 (8): 1-16. doi.org/10.1111/phc3.12760

Flores, A. W., K. Bechtel and C. T. 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).

Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., & Roth, D. (2019). “A comparative study of fairness-enhancing interventions in machine learning.” *Proceedings of the Conference on Fairness, Accountability, and Transparency*, Atlanta, GA, USA. https://doi.org/10.1145/3287560.3287589

Grgic-Hlaca, N., M. Bilal Zafar, K. P. Gummadi and A. Weller (2018). “Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning.” *Association for the Advancement of Artificial Intelligence*.

Hacker, P. (2018). “Teaching fairness to artificial intelligence: Existing and novel strategies against algorithmic discrimination under EU law.” *Common Market Law Review:* 1143-1185. Retrieved from http://www.kluwerlawonline.com/document.php?id=COLA2018095

Hedden, B. (2021). “On Statistical Criteria of Algorithmic Fairness.” *Philosophy and Public Affairs* 49 (2): 209-231.

Heidari, H., Ferrari, C., Gummadi, K. P., & Krause, A. (2019). “Fairness Behind a Veil of Ignorance: A Welfare Analysis for Automated Decision Making.” *arXiv e-prints*. Retrieved from https://arxiv.org/pdf/1806.04959.pdf

Hellman, D. (2008). *When Is Discrimination Wrong?* Cambridge: Harvard University Press.

Hellman, D., & Moreau, S. (Eds.). (2013). *Philosophical Foundations of Discrimination Law*. Oxford: Oxford University Press.

Hellman, D. (2020). “Measuring Algorithmic Fairness.” *Virginia Law Review,* 106 (4): 811-866.

Huq, A. Z. (2019). “Racial Equity in Algorithmic Criminal Justice.” *Duke Law Journal* 68: 1043-1134.

Kang, J., M. Bennett, D. Carbado, P. Casey, N. Dasgupta, D. Faigman, R. Godsil, A. G. Greenwald, J. Levinson and J. Mnookin. (2012). "Implicit Bias in the Courtroom." *UCLA Law Review* 59: 1124-1186

Khaitan, T. (2015). *A Theory of Discrimination Law*. Oxford: Oxford University Press.

Kleinberg, J., S. Mullainathan and M. Raghavan (2016). "Inherent Trade-Offs in the Fair Determination of Risk Scores." *arXiv e-prints*.

Kleinberg, J., Lakkaraju, H., Leskovej, J., Ludwig, J. & Mullainathan, S. (2018). “Human Decisions and Machine Predictions.” *The Quarterly Journal of Economics* 133 (1): 237-293.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Sunstein, C. R. (2019). “Discrimination in the Age of Algorithms.” *arXiv e-prints*.

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016). “How We Analyzed the COMPAS Recidivism Algorithm.” *ProPublica*. Retrieved from https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Lin, Z. J., Jung, J., Goel, S., & Skeem, J. (2020). “The limits of human predictions of recidivism.” *Science advances* 6 (7), doi:10.1126/sciadv.aaz0652.

Lippert-Rasmussen, K. (2006). “The Badness of Discrimination.” *Ethical Theory and Moral Practice* 9: 167-185.

Lippert-Rasmussen, K. (2013). *Born Free and Equal? A Philosophical Inquiry Into the Nature of Discrimination*. Oxford: Oxford University Press.

Lippert-Rasmussen, K. (2014). "Indirect Discrimination is Not Necessarily Unjust." *Journal of Practical Ethics* 2 (2): 33-57.

Lippert-Rasmussen, K. (2018a). “Respect and Discrimination.” In H. M. Hurd (Ed.), *Moral Puzzles and Legal Perplexities: Essays on the Influence of Larry Alexander:* 317-332. Cambridge University Press.

Lippert-Rasmussen, K. (Ed.) (2018b). *The Routledge Handbook of the Ethics of Discrimination*. Abingdon: Routledge.

Lippert-Rasmussen, K. (2020). *Making Sense of Affirmative Action*: Oxford University Press, Incorporated.

Lippert-Rasmussen, Kasper (2022), “Algorithm-based sentencing and discrimination.” In: Ryberg, J. & Roberts, J. (Eds.) *Sentencing and Artificial Intelligence*. Oxford: Oxford University Press.

Lipton, Z. C., A. Chouldechova and J. McAuley. 2018. “Does mitigating ML’simpact disparity require treatment disparity?” *32nd Conference on Neural Information Processing Systems*.

Liu, J. Z. and X. Li. 2019. "Legal Techniques for Rationalizing Biased Judicial Decisions: Evidence from Experiments with Real Judges." *Journal of Empirical Legal Studies* 16 (3): 630-670.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). “Algorithmic Fairness: Choices, Assumptions, and Definitions.” *Annual Review of Statistics and Its Application* 8 (1): 141-163. doi:10.1146/annurev-statistics-042720-125902

Moreau, S. (2020). *Faces of Inequality: A Theory of Wrongful Discrimination*: Oxford University Press, Incorporated.

Rachlinski, J. J., S. Johnson, A. J. Wistrich and C. Guthrie. 2009. "Does Unconscious Racial Bias Affect Trial Judges?" *Cornell Law Faculty Publications Paper* 786

Scanlon, T. (2008). *Moral Dimensions: Permissibility, Meaning, and Blame*. Cambridge: Belknap Press.

Shin, P. (2009). “The Substantive Principle of Equal Treatment.” *Legal Theory* 15 (2): 149-172.

Slavny, A., & Parr, T. (2015). “Harmless Discrimination.” *Legal Theory* 21(2): 100-114.

Thomsen, F. K. (2013). “But Some Groups Are More Equal Than Others - A Critical Review of the Group Criterion in the Concept of Discrimination.” *Social Theory and Practice* 39(1): 120-146.

Thomsen, F. K. (2015). “Stealing Bread and Sleeping Beneath Bridges - Indirect Discrimination as Disadvantageous Equal Treatment.” *Moral Philosophy and Politics* 2(2): 299-327.

Thomsen, F. K. (2022). “Iudicium ex Machinae - The Ethical Challenges of Automated Decision-Making at Sentencing.” In: J. Ryberg & J. V. Roberts (Eds.), *Principled Sentencing and Artificial Intelligence*. Oxford: Oxford University Press.

Thomsen, F. K. (2023). “No Disrespect – But That Account Does Not Explain the What is Morally Bad About Discrimination.” *Journal of Ethics and Social Philosophy* 23(3): 1-28. https://doi.org/10.26556/jesp.v23i3.1900.

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

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2019). “Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?” *Philosophy & Technology* 32: 661-683. doi:10.1007/s13347-018-0330-6.

Zimmermann, A., & Lee-Stronach, C. (2021). “Proceed with Caution.” *Canadian Journal of Philosophy*: 1-20. https://doi.org/10.1017/can.2021.17.

Wertheimer, A. (1983). “Jobs, qualifications and preferences.” *Ethics* 94: 99–112.

---

[i] I have presented draft versions of this article at research seminars with the Criminal Justice Ethics research group, Roskilde University, The Uehiro Centre for Practical Ethics, Oxford University, and the Centre for Experimental-Philosophical Study of Discrimination (CEPDISC), Aarhus University. I am grateful for valuable comments and suggestions on these occasions from Katrien Devolder, Thomas Douglas, Binesh Hass, Sebastian Jon Holmen, Rune Klingenberg Hansen, Ji Yong Lee, Kasper Lippert-Rasmussen, Søren Flinch Midtgaard, Lauritz Munch, Viki Pedersen, Jesper Ryberg, Aksel Sterri, Thomas Søbirk Petersen, and Søren Sofus Wichmann. I also owe thanks to three anonymous reviewers for the journal for persistent, lengthy and highly motivated comments over three rounds of review. Their objections helped me clarify my thinking on the issues even where we disagreed.