Algorithmic Indirect Discrimination, Fairness and Harm

Frej Klem Thomsen, fkt@dketik.dk, Danish National Centre for Ethics

Abstract: *Over the past decade, scholars, institutions, and activists have voiced strong concerns about the potential of automated decision systems to indirectly discriminate against vulnerable groups. This article analyses the ethics of algorithmic indirect discrimination, and argues that we can explain what is morally bad about such discrimination by reference to the fact that it causes harm. The article first sketches certain elements of the technical and conceptual background, including definitions of direct and indirect algorithmic differential treatment. It next introduces three prominent accounts of fairness as potential explanations if the badness of algorithmic indirect discrimination, but argues that all three are vulnerable to powerful levelling-down-style objections. Instead, the article demonstrates how proper attention to the way differences in decision-scenarios affect the distribution of harms can help us account for intuitions in prominent cases. Finally, the article considers a potential objection based on the fact that certain forms of algorithmic indirect discrimination appear to distribute rather than cause harm, and notes that we can explain how such distributions cause harm by attending to differences in individual and group vulnerability.*

Over the past decade, it has become increasingly more common that important decisions about issues such as banking, health care, social services, and criminal justice are made, not by a human decision maker, but by an automated decision-system (ADS) employing algorithmic profiling. At a very general level, such systems work as follows: a software algorithm is fed data about you, performs a statistical analysis of these data to calculate the probability that you do or do not possess a target property, and renders or recommends a decision based on whether that probability is above or below a specific threshold.

ADS potentially offers more efficient and precise evaluation of your case than a human assessment because it can rely on vast amounts of data, apply mathematically sophisticated analyses precisely and consistently, and avoid human cognitive biases. (M. Altman, Wood, & Vayena, 2018; Kleinberg, Lakkaraju, Leskovec, Ludwig, & Mullainathan, 2017; Zerilli, Knott, Maclaurin, & Gavaghan, 2019) The advent of ADS has not, however, been greeted with unanimous enthusiasm. Scholars, institutions, and activists have voiced strong concerns about the use of such systems. (e.g. AccessNow, 2018; Eubanks, 2018; European Group on Ethics

in Science and New Technologies, 2018; FRA, 2018; High-Level Expert Group on Artificial Intelligence, 2019; Jaume-Palasí & Spielkamp, 2017; Mittelstadt, Allo, Taddeo, Wachter, & Floridi, 2016; MSI-AUT, 2018; MSI-NET, 2017; O'Neil, 2016; Panel for the Future of Science and Technology, 2019; Rainie & Anderson, 2017; Reisman, Schultz, Crawford, & Whittaker, 2018; Zuiderveen Borgesius, 2018) One of the most prominent such concerns is that ADS can discriminate against vulnerable groups even when discrimination is neither intended nor coded into the ADS. I shall refer to this as indrect algorithmic discrimination (and clarify shortly).

Despite the critical attention it has drawn in public debate, there remains very little focused analysis of the ethics of indirect algorithmic discrimination within moral philosophy.[1] The central ambition of this paper is to remedy this situation by exploring indirect algorithmic discrimination from a philosophical perspective, and advancing a plausible account of what makes algorithmic indirect differential treatment morally bad, when it is.

The account I will defend holds that indirect algorithmic discrimination is morally bad when and because it causes harm. Perhaps there are further reasons why indirect discrimination is morally bad, but the harm-based account is both very plausible and capable of explaining our intuitions about important cases.

The argument proceeds as follows: Section 2 presents a conception of discrimination, particularly as regards the distinction between direct and indirect discrimination. Section 3 introduces demography, accuracy and error type as measures of indirect differential treatment and three corresponding parity conditions that have been widely discussed in the literature on fairness in machine learning. Section 4 presents a series of levelling down-type objections to argue that none of the parity conditions are plausible accounts of what is morally bad about indirect algorithmic discrimination. Section 5 introduces the harm-based account and notes how it can explain our intuitions once we take proper account of important differences between decision scenarios. Section 6 presents a challenge to the harm-based account, that it might be unable to explain the

[1] Some exceptions include (Binns, 2018; Hedden, 2021) [redacted]. Legal scholars have engaged more extensively with the issue, but with the emphasis and focus defined by that discipline. (See e.g. Chiao, 2019; Donohue, 2019; Hellman, 2020; Huq, 2019; Kleinberg, Ludwig, Mullainathan, & Sunstein, 2019; Roth, 2016; Zarsky, 2016; Barocas & Selbst, 2016)

moral badness of scenarios where harmful classifications are unavoidably distributed. Section 7 answers the challenge and further details the harm-based account by pointing out how differences in individual and group vulnerabilities affect the harms produced by classification. Jointly, sections five to seven demonstrate that we can explain the moral badness of indirect algorithmic discrimination in relevant cases by reference to the fact that the ADS causes harm. Section 8 concludes by briefly considering some implications and limitations of the argument.

Let me make two notes of clarification before we begin. I shall speak throughout of discrimination being "morally bad". By something being morally bad, I mean that it is a wrong-making feature of the act, which is to say that it grounds a *pro tanto* reason against the act. I thus leave open the possibility that there could be situations morally bad discrimination is all-things-considered permissible.

Furthermore, I shall speak throughout of "majority groups" and "majority persons" versus "minority groups" and "minority persons". I use these terms as placeholders for the groups we might be concerned with in the context of algorithmic indirect discrimination. In practice it will often be important which specific groups are at stake – I discuss two reasons why in section 7 – but for many general points it will be helpful to consider differential treatment of persons and groups in the abstract.

2. What is algorithmic indirect discrimination?

The focus of this article is the moral badness of indirect algorithmic discrimination. One difficulty for the analysis is that discrimination is a complex phenomenon.[2] In this section, I define and distinguish between direct and indirect discrimination, and briefly sketch how algorithmic indirect differential treatment can occur.

[2] Over the past decade (roughly) there has been increased philosophical interest in discrimination, which the current analysis draws on. Prominent contributions include (Collins & Khaitan, 2018; Eidelson, 2015; Hellman, 2008; Hellman & Moreau, 2013; Khaitan, 2015; Lippert-Rasmussen, 2013, 2018b, 2020; Moreau, 2020)

What are we to understand by "discrimination", then? I adopt here a moralised definition of discrimination, such that an act is discrimination only if it is in a relevant respect morally bad.[3] Drawing on, but simplifying, an influential analysis by Kasper Lippert-Rasmussen, I shall say that:

> **Direct discrimination.** An agent *directly* discriminates against minority persons *iff* (i) she treats minority persons differently than she treats or would treat majority persons, (ii) her treatment of minority persons is disadvantageous, (iii) the difference in treatment is caused by minority persons being minority persons, and (iv) the differential treatment is in some respect morally bad. (Cf. Lippert-Rasmussen, 2013, 2006)

On this account, an ADS directly differentially treats a group if it employs membership of that group as a variable in its function, such that prediction of the target property depends on the value of the variable (member/non-member). For example, an ADS that increases the predicted market value of a house if the seller is male directly treats men and women differently. If the differential treatment is disadvantageous to women and morally bad, then the ADS directly discriminates against women.

The crucial distinction for our purposes is between direct and indirect differential treatment. The conditions that make the above a definition of *direct* discrimination, which involves direct differential treatment, are (i), and (iii). How ought we to modify them, to define indirect discrimination? I want to say that:

> **Indirect discrimination.** An agent *indirectly* discriminates against minority persons *iff* (i) she treats (in the acts at stake) minority persons *equally* to how she treats or would treat majority persons, (ii) the equal treatment is disadvantageous to minority persons, (iii) the

[3] I am not persuaded that this is generally speaking the best way to define discrimination, but for present purposes it will make the discussion easier. (cf. Eidelson, 2015; Lippert-Rasmussen, 2013) [self-reference removed for purposes of anonymity] Note also that I do not restrict discrimination to a particular set of groups. Again, I do not think that employing a group criterion is the best way of conceiving of discrimination [self-reference removed for purposes of anonymity], but my choice to exclude it here is strictly pragmatic: we are using minority person/majority person as placeholders to explore certain analytical points, and the analysis is compatible with both a generic and a narrow conception of the groups that can be filled in.

> difference in *effect* is caused by minority persons being minority persons, and (iv) the treatment is in some respect morally bad. [4]

The notion of equal treatment at stake here is purely formal. An ADS treats two groups equally in this sense so long as its function does not employ a variable that corresponds to group membership to predict the target property. Thus, an ADS *indirectly* differently treats a group if it employs at least one variable that correlates with membership of the group, such that it has a tendency to predict different values for members of the group. For example, an ADS that employs "height" as a variable will thereby indirectly treat men and women differently because height correlates with gender.

It is also important to bear in mind, that an ADS can treat groups differently (indirectly) both when the correlations at stake are and are not spurious. (Corbett-Davies & Goel, 2018; Mitchell, Potash, Barocas, D'Amour, & Lum, 2018) In the former case, the ADS can be trained on biased data, in which case it will tend towards treating groups differently because it assigns inordinate weight to the relevant variable(s), or it can be fed biased data when employed, in which case even an ADS that assigns the appropriate weights will differentially treat groups. A common example is the uneven registration of criminal activity caused by discriminatory police patrolling, which may inflate the perceived criminality of over-patrolled neighbourhoods and their residents. (Ensign, Friedler, Neville, Scheidegger, & Venkatasubramanian, 2017; Ferguson, 2017; Perry, 2013)

An ADS can also indirectly differently treats groups based on *accurate* correlations between variables, groups and targets. If, to continue our (hopefully) innocent example from above, height correlates with both prowess in the sport of basketball and gender, then an ADS that attempts to predict basketball prowess and employs height as a variable will indirectly differentially treat men and women, because it tends to (accurately) predict

[4] [Self-reference removed for purposes of anonymity] (cf. A. Altman, 2020; Cosette-Lefebvre, 2020; Eidelson, 2015; Lippert-Rasmussen, 2013; Khaitan, 2017) Regardless of what one thinks is the best general account of indirect discrimination, the definition I adopt here seems to me very useful as a way of drawing out the distinction between direct and indirect differential treatment in ADS.

higher scores for men than for women. In many less innocent cases, such correlations may themselves be the result of both historical and ongoing discrimination of these groups.

Finally, it is worth noting that ADS often employ dozens, hundreds or even thousands of variables. Furthermore, these correlations can exist without developers being aware of them, such that indirect differential treatment can and likely often will occur inadvertently. In light of these facts, we urgently need a way to identify those cases where an ADS is indirectly discriminatory, which is to say we need an account of what makes algorithmic indirect differential treatment morally bad (when it is). It is to this issue, that we now turn.

5. What might be bad about algorithmic indirect discrimination?

I have suggested above that an ADS indirectly differentially treats groups when it employs variables that correlate both with the target property and group membership. The result will be systematic differences in outcomes across groups. When does this amount to discrimination? That is, when is such indirect differential treatment morally bad?

Discussion of this issue within the data science and machine learning literature has mostly been conducted under the heading of "fairness". (Barocas, Hardt, & Narayanan, 2019; Berk, Heidari, Jabbari, Kearns, & Roth, 2018; Chouldechova & Roth, 2018; Dwork, Hardt, Pitassi, Reingold, & Zemel, 2011; Hacker, 2018; Kleinberg, Mullainathan, & Raghavan, 2016) "Fairness", however, appears to often be used as a placeholder for ethical concerns more broadly. Thus, I shall assume that we are interested in explaining the moral badness of algorithmic indirect discrimination, regardless of whether this explanation happens to turn on a philosophically precise conception of fairness.[5]

Scholars within the data science and machine learning community have developed a range of accounts of what it means for an ADS to be fair. (Berk et al., 2018; Chouldechova & Roth, 2018; Friedler et al., 2019;

[5] There are theoretically precise conceptions of fairness in the philosophical literature. Prominent examples include (Broome, 1990; Daniels, 2008; Otsuka & Voorhoeve, 2009; Rawls, 1999). See also (Binns, 2018).

Hedden, 2021; Mitchell et al., 2018; see Mitchell, Potash, Barocas, D'Amour, & Lum, 2021 for an excellent overview) At a general level, the most important differences between these accounts boil down to three measures of affecting groups differently: demography, accuracy, and error type.

Demography measures the ADS' distribution of positive and negative classifications.[6]

Differential effect in terms of demography means unequal rates of positive to negative classifications. To measure demography, we can ask: what is the probability that a randomly selected person from the relevant group will be positively classified? Subsequently, we can compare these scores for minority persons and majority persons. If the scores are significantly different, then the ADS differentially treats the two groups with respect to demography.[7]

Accuracy measures the ADS' distribution of true and false **classifications**.

Differential effect in terms of accuracy means unequal rates of true to false classifications. To measure accuracy, we can ask: what is the probability that a randomly selected person from the relevant group is correctly classified? Subsequently, we can compare these scores for minority persons and majority persons. If the scores are significantly different, then the ADS differentially treats the two groups with respect to accuracy.[8]

[6] I shall discuss mostly classification problems, where ADS attempts to predict the presence of a target property ("Is this person a woman?"), but the points transfer readily to regression problems, where the ADS attempts to predict the value of the target property ("What is this person's age?").

[7] "Significantly" because we must allow for minor differences attributable to randomness. Also, note that negative and positive classifications are symmetrical in the sense that increasing the probability of one equally reduces the probability of the other. Thus, we need only review one of the two to measure disparity.

[8] One can further specify accuracy for specific sub-groups. We can, for example, measure accuracy for those positively/negatively classified by the model (precision or positively predictive value, and negative predictive value, respectively) or for those who do or do not in fact possess the target property (sensitivity/recall/true positive rate, and specificity/selectivity/true negative rate, respectively). (Mitchell et al., 2018) These specifications will often be important in practice. For example, predictive value is more useful when a decision-maker needs to determine how much evidentiary weight to attribute to the model's prediction in a specific case, while true rates are more useful as an estimate of the odds a specific person has of receiving a correct prediction. (cf. Hellman, 2020) Much of the discussion around fairness has focused on these more specific accuracy measurements, including a number of celebrated proofs that in realistic scenarios models cannot simultaneously achieve parity in more than a subset of specific accuracy measurements. (Kleinberg et al., 2016; Chouldechova, 2017; for discussion, see Berk et al., 2018; Hedden, 2021; Heidari, Ferrari, Gummadi, & Krause, 2019) However, all of these measures ultimately concern the distribution of true/false

> **Error type** measures the ADS' distribution of false positive and false negative errors.

Differential effect in terms of error type means unequal rates of false positives to false negatives. To measure error type, we can ask: what is the probability that a randomly selected person from the relevant group *who is misclassified* receives a false positive? Subsequently, we can compare these scores for minority persons and majority persons. If the scores are significantly different, then the ADS differentially treats the two groups with respect to error type.

Now, for each of the three measures, we might require that ADS treats minority persons and majority persons equally in the pertinent sense. Thus, we might require:

> **Parity of demography (PD):** Demography of the minority group is (roughly) equal to demography of the majority group.[9]

Demographic parity ensures that classifications are distributed to minority persons in proportion to their fraction of the total population. Thus, if women are half the persons in a society, then PD requires that they will receive half of the positive and half of the negative classifications, while allowing, for example, that both minority and majority persons are more frequently classified negatively than positively. An often discussed application is representation in political and business leadership. Most persons will not obtain a position as a political or business leader, but we may consider it desirable that among the small group of persons who *do* become such leaders roughly half are men and roughly half women.

We might also require:

> **Parity of accuracy (PA):** Accuracy for the minority group is (roughly) equal to accuracy for the majority group.

predictions, and as such the levelling down objection I raise against parity of accuracy below applies equally to each of the potential more specific measurements.

[9] "Roughly" because, as before, we should presumably allow space for minor differences attributable to randomness.

Parity of accuracy ensures that the ADS will be equally good at avoiding mistakes when classifying persons from the two groups. Thus, if Black, Asian and Middle-Eastern (BAME) persons make up one third of society, PA requires that (roughly) a third of the errors the algorithm makes affect BAME persons. We might for example require that an ADS employed by a public child protection service to flag children in certain families as at risk, does not more or less frequently make mistakes when reviewing BAME families than when it reviews majority families.

Finally, we might require:

> **Parity of error type (PET):** Error type for the minority group is (roughly) equal to error type for the majority group.

Parity of error type ensures that the ADS will be equally likely to make the two kinds of mistake when classifying persons from the two groups.[10] Thus, if false positives make up half the errors when assessing theists, then PET requires that false positives make up (roughly) half the errors when assessing atheists. Notably, this is compatible with the ADS having different accuracy for the two groups – PET requires only that the ratios be equal for the errors actually made. In the most widely discussed example of applying this principle, if the criminal justice system relies on evaluations of offender risk of recidivism to determine bail or parole, then it makes a big difference to an offender whether a mistake is an over- or underestimation of her risk. (Cf. Angwin, Larson, Mattu, & Kirchner, 2016; Dressel & Farid, 2018; Dieterich, Mendoza, & Brennan, 2016; Larson, Mattu, Kirchner, & Angwin, 2016)

In the remainder of this article, I will advance two central arguments. First, that failure to meet any or all of the parity-conditions does not by itself make algorithmic indirect differential treatment morally bad, and

[10] As with accuracy above, we can specify more specific conditions, such as parity of false positives. We could also require parity of true type, that is, that the ADS is equally likely to make the two kinds of accurate assessment when classifying persons from the two groups. This requirement is rarely suggested, however, perhaps because we tend to focus on socially beneficial ADS, where accurate assessments are assumed to be justified. Regardless, the problems afflicting PET, discussed below, apply equally to the related "true type" or more specific parity conditions.

second, that the harm such treatment can cause *does* make algorithmic indirect differential treatment morally bad, and is thus a plausible explanation of indirect algorithmic discrimination.

6. Against algorithmic parity

Much of the discussion of the moral badness of indirect algorithmic discrimination has focused on the relative advantages and disadvantages of the parity conditions above and their various specifications. Famously, it has been proven that it is in realistic circumstances impossible to simultaneously satisfy certain sets of parity conditions, forcing developers to choose how to prioritize them. (Chouldechova, 2017; Kleinberg et al., 2016) Some believe that this gives rise to a trilemma of roughly the following form:

1) ADS that does not satisfy parity conditions [P_1, P_2… P_n] is morally bad algorithmic indirect differential treatment.
2) ADS cannot (in real-life cases) simultaneously satisfy [P_1, P_2… P_n].
3) ADS is not (in real-life cases) unavoidably morally bad algorithmic indirect differential treatment.

The three claims are logically incompatible – at least one of them is false. As noted above, the second claim has been mathematically proven, and the third is a very plausible ethical claim. Hence, the first claim must be false. Much of the debate has thus focused on figuring out which parity conditions to preserve and which to abandon. In this section, however, I will argue that PD, PA and PET are each vulnerable to powerful levelling down-type objections.[11] The answer to the question of which parity condition we must satisfy in order to avoid indirect algorithmic discrimination is "none".

Let us consider first so-called parity of demography. PD, recall, obtains simply when the ratio of minority persons who receive a positive classification relative to the number of minority persons in the population is

[11] The original levelling down objection was famously raised by Derek Parfit against telic egalitarianism. (Parfit, 2002) I say levelling down-*type* objections, because they are structurally similar but different in that they pertain to reasons, as opposed to values. (cf. Lippert-Rasmussen, 2015, chapter 5)

(roughly) equal to the ratio of majority persons who receive a positive classification relative to the number of majority persons in the population. Consider now:

> **Medical**. A hospital employs ADS to determine whether a particular form of cancer is malignant or benign. If untreated, malignant tumours are fatal, but treatment also has very bad side-effects.

In *Medical*, PD obtains between men and women if the proportion of female patients who are diagnosed as having a malignant tumour is equal to the proportion of male patients who are diagnosed as having a malignant tumour. If, for example, one in five female patients is diagnosed as having a malignant tumour, then demographic parity between men and women obtains if one in five male patients are also diagnosed as having a malignant tumour. The question now is this: is there a moral reason to strive for demographic parity in ADS?

Although we often discuss PD as an important policy goal, e.g. in the context of representation of minority persons in political or business leadership positions, it seems to me clear that there is no moral reason to strive for PD in and of itself. Consider:

> **Medical 2**. A hospital employs ADS to determine whether a particular form of cancer is malignant or benign. If untreated, malignant tumours are fatal, but treatment also has very bad side-effects. In majority persons, tumours are 50% likely to be benign and 50% likely to be malignant. In minority persons, tumours are 80% likely to be benign and 20% likely to be malignant. The two groups are equally big.

In *Medical 2*, ADS with a 100% accuracy will assess far more majority persons than minority persons as having a malignant tumour (1:2 vs 1:5). This means that PD does not obtain (1:2 >> 1:5).[12] It would be strange to say, however, that this disparity is morally problematic. After all, it simply reflects the fact that tumours are more

[12] The same point applies to benign tumours, of course, though for simplicity we can focus on only one of the two.

likely to be malignant in majority persons than in minority persons. In this case, we could achieve (or approach) PD only by misdiagnosing persons. This would be bad for these persons regardless of whether they were minority persons with malignant tumours misdiagnosed as healthy, or healthy majority persons misdiagnosed as in need of treatment. In either case, therefore, we would be "levelling down", in the sense that we would be promoting equality (i.e. PD) by making some persons worse off without making anyone (non-comparatively) better off. This levelling down would, it seems clear to me, not merely be morally bad, it would be a change that there is *no* moral reason to pursue. A plausible explanation of what has gone wrong is that PD fetishizes classifications, ignoring the way the value of a classification can vary from person to person and context to context.[13]

Could we avoid the problem by further specifying the groups? We might require, for example, that the proportion of *healthy* majority persons (i.e. those with a benign tumour) who receive a benign diagnosis is equal to the proportion of *healthy* minority persons who receive a benign diagnosis. While this would avoid the specific problem in *Medical 2* and focus on the apparently relevant distinction within the groups, all we have done is rediscover one of the specifications of parity of accuracy (which we consider below). Thus, the suggested focus on healthy persons is equivalent to a requirement that ADS have equal true positive rates for the two groups.

What then of parity of accuracy? PA obtains, as we noted above, when the probability of receiving a true classification is roughly the same for minority persons and majority persons. Consider:

> **Medical 3.** A hospital employs ADS to determine whether a particular form of cancer is malignant or benign. If untreated, malignant tumours are fatal, but treatment also has very bad side-effects. Accuracy for majority persons is 99%; accuracy for minority persons is 95%.

[13] For a related general argument, see (Lippert-Rasmussen, 2008).

In *Medical 3*, parity of accuracy does not obtain. Is this morally bad? If it were, then retraining the ADS to achieve or approach PA should in at least one respect be a moral improvement. We can suppose, however, that we cannot improve accuracy for minority persons, so that the only way to approach PA is to reduce accuracy for majority persons. As with PD above, this levelling down would, it seems clear to me, be a wholly awful thing to do since it would result in misdiagnoses of hapless majority persons while making no minority person better off. That cannot be a moral improvement in any respect, and as such, failure to satisfy PA cannot be morally bad.

Consider finally parity of error type. PET obtains, as we noted above, when the probability of a mistaken classification being a false positive (as opposed to a false negative) is (roughly) the same for minority persons and majority persons. PET has seemed appealing to many in the debate on fairness in ADS, but it is vulnerable to an objection very similar to those encountered by PD and PA above. Consider:

> **Medical 4.** A hospital employs ADS to distinguish benign from malignant tumours. If untreated, malignant tumours are fatal for majority persons, but treatment also has very bad side-effects. Untreated malignant tumours are non-fatal for minority persons, although still worse than treatment. The ADS has equal accuracy for the two groups, but makes more false positive errors for majority persons, and more false negative errors for minority persons.

Let us suppose that we could obtain parity of error type by shifting the decision boundary for majority persons only, and let us further suppose (less realistically) that this would not substantially affect accuracy for the group. That is, we retrain the ADS so as to increase the number of false negatives and decrease the number of false positives for majority persons, without changing the overall number of errors. Is there any reason to do so? It strikes me as absurd to suppose that the answer could be "yes", since false negatives are worse than false positives, false negatives are *even worse* for majority persons than false negatives are for minority persons, and no minority person would benefit in any way from the redistribution of errors within

the majority group. Once again, if there are situations where we have no reason to promote PET, then failure to satisfy it cannot be what makes indirect algorithmic discrimination morally bad.

It is worth noting that committed egalitarians have defended the view that levelling down can in some cases be better in *one respect*, to wit, in the respect that the outcome is more egalitarian. (Temkin, 2002; Otsuka & Voorhoeve, 2009; Voorhoeve & Fleurbaey, 2012) Does this offer friends of the parity conditions an avenue of response? Not a promising one. In part this is because, although the subject of a complex debate, it is not clear that egalitarian counters are persuasive. (Crisp, 2011; Holtug, 2010; Parfit, 2012) More importantly, theoretically developed defences of egalitarianism do not translate well to the context of parity conditions. Egalitarians are concerned with whole-life distributions of intrinsically valuable goods. One point illustrated by the *Medical* cases above is that demography, accuracy and error type do not correspond to and often do not correlate well with such goods. One way to illustrate this point is the ease with which we can imagine levelling down scenarios where approaching a parity condition requires increasing whole-life inequality. Egalitarian intuitions will support the parity conditions only when approaching parity promotes equality, e.g. when it makes a disadvantaged minority group better off or a privileged majority group worse off, not when it requires the converse. The parity conditions, however, are symmetrical, that is, they take the moral reasons in favour of the two types of levelling down to be equal. Thus, even committed egalitarians are likely to find the parity conditions unappealing.

I conclude that all three parity conditions are vulnerable to powerful levelling down-type objections, and as such it seems we will have to look elsewhere for an explanation of what makes indirect algorithmic discrimination morally bad.[14] In the next three sections of the article, I will argue that the harm-based account provides such an explanation.

[14] The levelling down-type objections apply even if the scenarios are unlikely to occur in practice, but it is worth noting that such scenarios may in fact be common. (Corbett-Davies, Pierson, Feller, Goel, & Huq, 2017; Corbett-Davies & Goel, 2018)

7. Classification scenarios and harm

Harm-based accounts of the badness of discrimination take point of departure in the claim that causing harm is morally bad. (Arneson, 2017; Lippert-Rasmussen, 2006, 2007, 2013) [self-reference removed for purposes of anonymity] This is perhaps the least controversial claim in moral philosophy – if anything is morally bad, then surely causing harm is – and as such, I will not defend that claim here. If we grant the assumption that causing harm is morally bad, then algorithmic indirect differential treatment that causes harm is morally bad, which is to say that it constitutes indirect algorithmic discrimination. Does indirect algorithmic differential treatment cause harm? Sometimes, yes. In this and the next two sections, I detail the different ways it can do so, focusing first on differences in the value of classifications across decision scenarios, and next on differences in vulnerability.

Perhaps the most obvious way ADS can cause harm is when the way it classifies a person deprives that person of a good or imposes upon them a bad. This simple observation can explain many of our intuitions about ADS being morally bad. For example, PD may seem appealing if we assume equal base rates in scenarios like *Medical*, because unequal demography then suggests that the ADS misclassifies persons, producing (more) false positives for one group, (more) false negatives for the other, or both. PD loses its attraction in *Medical 2*, where unequal base rates mean that there is no longer any correlation between failure to achieve PD and harmful misclassification. The same point applies to PA and PET in Medical 3 and Medical 4 respectively: while appealing in some scenarios, they lose their appeal when approaching or achieving them no longer correlates with decreasing harm.

There are some important complications to bear in mind, however. The first is that the ways ADS can be bad for a person varies in significant ways across different types of decision scenario. Some decision scenarios distribute goods across true and false classifications, while others distribute across positive and negative classifications. This difference both helps to explain intuitions about some prominent cases and to further illuminate how algorithmic indirect discrimination can cause harm.

In *Medical*, the distribution of goods is across true/false classifications. A false positive is an assessment that mistakenly labels a benign tumour as malignant. A false negative is an assessment that mistakenly labels a malignant tumour as benign. Thus, any *false* classification is bad for the person receiving it. Consider now for comparison:

> **Criminal**. A penal system grants parole to offenders based largely on predicted recidivism, and employs ADS to determine whether offenders are likely to reoffend upon receiving parole. Parole allows offenders to escape the dangerous and oppressive environment of prison to pursue the benefits and opportunities of civilian life.

In *Criminal*, a false positive mistakenly labels a low-risk offender as a high-risk recidivist. This is bad for the person because it increases the probability of being denied parole. A false negative mistakenly labels a high-risk offender as a low-risk recidivist. This is good for the person because it increases the probability of being granted parole.[15] Thus, in *Criminal*, the distribution of goods is across positive/negative classifications. Any *positive* classification is bad for the person.

These observations helps to explain intuitions in prominent cases. One of the standing concerns in the debate on fairness in ADS is how to prioritise between parity of accuracy and parity of error distribution. In some decision scenarios, the former looks more appealing while in others the latter seems to be more important. We can account for (at least some of) these apparently conflicting intuitions by referencing the difference between *Medical*-type scenarios and *Criminal*-type scenarios, specifically the fact that there is a difference with regards to which classifications are bad for persons.

[15] We set aside here the possibility that it may be all things considered worse for the person likely to reoffend to be granted parole, e.g. because this will allow them to reoffend, and reoffending is bad for the offender. Furthermore, we are still setting aside the issue of when an act, policy or practice might be all-things-considered permissible in spite of the fact that it is bad for some persons, e.g. because denying parole to persons accurately assessed as high-risk recidivists prevents harm to potential victims.

In *Medical*-type scenarios, persons evaluated by the ADS will want accuracy to be as high as possible, and lack of PA means that one group is more likely than the other to be classified in a way that is bad for persons. In *Criminal*-type scenarios, persons evaluated by the ADS will want error distribution to be tilted towards false negatives as much as possible, and lack of PET means that one group is more likely than the other to be mistakenly classified in the way that is bad for the person.

If ADS can classify persons in ways that are bad for them, have we successfully shown that harm can explain algorithmic indirect discrimination? Not quite, for not all algorithmic indirect differential treatment that assigns a classification that is bad for a person causes harm.

8. Unavoidable bad classifications and harm

What is the relation between classifications that are bad for a person and harm? Consider again *Medical 3,* where the ADS has 99% accuracy for majority persons but only 95% accuracy for minority persons. Because of the way the ADS is trained, it indirectly differentially treats minority persons, more often misclassifying them, and misclassifications are bad for persons. This might tempt us to jump immediately to the conclusion that the ADS' indirect differential treatment is morally bad because it causes harm. This would be a mistake, however, because there are cases where bad classifications are *unavoidable*, and in such cases the ADS' indirect differential treatment arguably does not cause harm.

Thus, in *Medical 3*, whether the ADS harms minority persons seems to depend on background conditions that determine *why* parity of accuracy does not obtain. Consider the following three possibilities: the ADS' accuracy for minority persons is i) pareto-suboptimal, ii) optimal, or iii) pareto-optimal.

The ADS' accuracy for minority persons is *pareto-suboptimal* if the ADS could be retrained with improved accuracy for the minority group and no loss of accuracy for the majority group. In this case, the ADS' indirect differential treatment clearly causes harm. There are persons, who will be badly off because they receive the wrong diagnosis, and this is avoidable simply through better training.

However, an algorithm can indirectly differentially treat two groups without the latter condition being met. Lower accuracy for one group may for example be an unavoidable consequence of limitations in the data the developers have for training. The accuracy of the algorithm is *optimal* with respect to minority persons if the ADS could not be retrained with improved accuracy for the minority group. In this case, although the ADS indirectly differentially treats minority persons, by more frequently misclassifying them, and misclassifications are bad for persons, it is *not* true that minority persons are harmed (in this way) by the indirect differential treatment, because there is no way to prevent these misclassifications.[16]

The implications of pareto-suboptimality and optimality are simple and relatively uncontroversial. Attention in arguments on fairness and the parity conditions has often been justifiably focused on the third possibility, where the accuracy of the ADS is *pareto-optimal* (but not optimal) with respect to minority persons. That is, the ADS could be retrained with improved accuracy for the minority group, but only at the loss of some accuracy for the majority group. This scenario might occur e.g. if the predictive value of variables differs across groups, and the total accuracy of the ADS is optimised by drawing the decision boundary in a way that more commonly misclassifies minority persons. In some cases, increasing accuracy for one group will lower accuracy for the other group to an even greater extent. In other cases, it may be possible to train the ADS in slightly different ways, that all have the same overall accuracy but redistributes mistakes across groups. Such cases might be rare for certain decision scenarios, but common in others. Consider:

> **Educational**. A university accepts students based on expected academic performance, and employs ADS to rank applicants. There are hundreds of applicants but only the 100 highest ranked applicants are admitted. Education is both a means of valuable self-development and a qualification for many attractive positions on the labour market.

[16] Does it matter what the alternatives to ADS are in the first place, for example how a human doing the same classification task would perform? Yes, clearly. The ADS causes harm if we could do better without it. (Cf. M. Altman et al., 2018) For the purposes of this argument, however, such alternatives ("the human ADS") are no different than the possibility of training a different model. Hence, let us assume that alternatives to ADS are impossible or would be even worse.

In *Educational*, overestimation of an applicant's academic potential, for example being ranked 71st when one's academic potential really only merits a score of 82nd, is good for the person (in expectation, at least).[17] Underestimation of an applicant's academic potential is bad for the person for similar reasons. As in *Criminal*, the valence of overestimation (false positives) and underestimation (false negative) differs, since any low rank (negative classification) is bad for the person. However, in *Medical* and *Criminal*, the distribution of goods to any individual is independent of how other persons are classified. In *Educational*, predicted potential is a *positional* good: there are a limited number of absolute goods (education), which are distributed on the basis of the ADS' ranking.

This has two important implications. First, the ADS' treatment of each person directly affects others. Because ranking is a positional good, it can be good for a person that *other* persons receive low rankings (negative classifications), and bad for a person that other persons receive high rankings (positive classifications). Second, the ADS' predictions do not affect the amount of absolute goods distributed: there are exactly 100 positions, and the number of applicants refused is determined entirely by the number of total applicants. As such, training of the ADS cannot affect the *number* of predictions that are bad for a person, only the *identity* of the persons receiving these.

In at least some cases it may therefore simultaneously be true that some persons are worse off because of the way the ADS classifies them, and false that the ADS or its indirect differential treatment is bad because it causes harm. The former because we could retrain the ADS so as to bring it about that some persons received a classification that is good for them; the latter because retraining the ADS would not lead to fewer bad

[17] Overestimation is only *actually* good when it makes a difference to whether the person obtains an education or not. We set aside for simplicity's sake the complex issue of what it means to have academic potential, and whether it can plausibly be ranked. That is, we assume for the purposes of the argument that we can meaningfully speak of a rank that one really merits. Furthermore, as in *Criminal*, we set aside here the possibility that some persons may be *worse* off by being overestimated, e.g. because they are offered and accept a position at an education they are incapable of completing, and the resulting waste of time and experience of failure leave them worse off than they would have been, had they not been offered a position at all.

classifications, and so to less harm, but only change the identity of the persons who receive a classification that is bad for them.

This might appear to have counterintuitive implications for the harm-based account of indirect algorithmic discrimination. In such situations, it might seem, we could not condemn indirect differential treatment as discriminatory, even if the group of those persons unavoidably harmed is disproportionately composed of minority persons. Let us call this objection:

> **Musical chairs:** In some cases, algorithmic indirect differential treatment only changes the *identity* of the persons who receive a classification that is bad for a person; it does not change the *number* of persons who receive a classification that is bad for a person. In these cases, the ADS is not bad because it causes harm. But some of these cases are indirect algorithmic discrimination. As such, harm cannot (at least, by itself) explain what is morally bad about algorithmic indirect discrimination.[18]

The possibility of pareto-optimal indirect differential treatment leaves the harm-based account of indirect algorithmic discrimination with an explanatory challenge. In the next and penultimate section, I will argue that we can further elaborate on the harm-based account to explain why it often matters how classifications that are good and bad for a person are distributed across minority and majority persons.

9. Individual and group vulnerabilities

The central argument for the harm-based account of algorithmic indirect discrimination so far has been to show how the value of classifications can vary between decision scenarios. In addition to this, however, the value of classifications can vary across the persons and groups affected. Some persons and groups are *vulnerable*, and this vulnerability means that they will be more adversely affected by indirect differential

[18] Note that the objection does not purport to show that harm explains the badness of *no* cases of indirect algorithmic discrimination. In fact, it is compatible with the objection that harm explains the badness of many cases. The objection is an argument for the more modest claim that harm cannot explain the badness of *all* cases, and that there must therefore be more moral factors at stake.

treatment. I will argue in this section that awareness of such differences in vulnerability has important implications for how we should understand indirect algorithmic discrimination, and that it allows the harm-based account to answer the *Musical chairs* objection, by explaining why indirect differential treatment of some persons and groups matters more than indirect differential treatment of others.

In the above, I have argued that one important reason why indirect algorithmic differential treatment can be morally bad is that it may cause harm. However, we have implicitly assumed (for the most part) that the value or disvalue of a classification is the same for all persons.[19] This assumption is likely to be false in many cases. Persons will differ in their vulnerability, understood as the degree to which they are harmed by receiving a classification that is bad for a person. Such individual vulnerability will in many cases correlate with membership of minority and majority groups. (Cf. M. Altman et al., 2018) We encountered a toy example of this in *Medical 4*, where malignant tumours affected one group more severely than the other, making false negatives worse for majority persons than for minority persons. Minority/majority group correlated differences in vulnerability occur in many real-life scenarios because of the correlation between groups and socio-economic status on the one hand, and socio-economic status and vulnerability on the other.

This simple observation goes a long way towards answering the *Musical chairs*-objection. The persons and groups we tend to focus on in the context of indirect discrimination are typically vulnerable, in many cases because of past marginalisation and discrimination. This vulnerability means that minority persons tend to suffer greater harm than majority persons from receiving the same (bad) classification. This in turn explains why an ADS may cause harm by distributing classifications that are bad for a person from majority to minority persons, even when there is no change in the number of classifications that are bad for a person.[20] Developers who wish to avoid such harm must take account of individual vulnerability. It seems clear that we ought to do so in *Medical 4* – it is more important to avoid the false negatives that kill patients, than the

[19] As (Mitchell et al., 2018) observe, this dubious assumption is common in both development of ADS and academic discussions of fairness in machine learning.

[20] It is also possible, as prioritarianism claims, that the moral value of units of wellbeing vary with the well-being level of the recipient. If so, harming persons who are in general worse off is morally even more bad.

false negatives that only make patients very ill (bad as that may be) – but exactly the same point applies across all of the contexts where persons differ in vulnerability.

What about group vulnerability? The vulnerabilities I have discussed above remain individual vulnerabilities, even if they correlate with membership of minority groups. Furthermore, there is a sense in which the harms we have been discussing are not really harms that occur *because* of discrimination. It is not the fact of *differential* treatment that causes harm, but the bad classifications and individual vulnerabilities by themselves, as evinced by the fact that in cases like *Medical 3* we could achieve equal treatment by reducing accuracy for majority persons without thereby reducing the harm minority persons suffer.

There are harms, however, that can follow specifically from differential treatment. The most familiar examples of these occur in cases of direct discrimination. Consider:

> **Medical 5.** A hospital performs diagnoses to determine whether a particular form of cancer is malignant or benign. The hospital employs ADS when diagnosing majority persons, and coin-flipping when diagnosing minority persons. If untreated, malignant tumours are fatal, but treatment also has very bad side-effects.

The diagnostic procedure for minority persons – flipping a coin – will of course produce many more errors than the ADS used for majority persons (if not, the hospital urgently needs to retrain their ADS). This is at least one reason why the direct differential treatment in *Medical 5* is bad (i.e. discrimination). However, the differential treatment may also cause harm in a different way. If the direct differential treatment becomes public knowledge, then minority persons will reasonably feel neglected and humiliated. Some persons, majority as well as minority, may also mistakenly take the differential treatment as evidence that the two groups differ in morally relevant respects, or as support for their negative attitudes towards members of the group. Stereotyping, stigmatisation, loss of trust and cooperation, loss of self-esteem, resentment, and discrimination in other contexts are realistic consequences of the awareness of the differential treatment. (See e.g. Benner et al., 2018; Berger & Sarnyai, 2015; Krieger, 2014; Schmitt, Branscombe, Postmes, & Garcia,

2014; Williams, Lawrence, Davis, & Vu, 2019; Cf. also Huq, 2019) Plausibly these effects constitute harms, and as such are harms that follow specifically from the differential treatment. Notably, they are harms that occur because of group vulnerabilities, and which apply to members of groups beyond the individual classified. While these harms are more familiar in the context of direct differential treatment, there is no reason to think that they cannot be caused by indirect differential treatment.

Group vulnerabilities are the second and final piece of the puzzle about the ways even pareto-optimal indirect differential treatment can cause harm. In combination with awareness of the effect of individual vulnerabilities, we can answer the *Musical chairs*-objection, and in combination with our understanding of differences in decision scenarios, we can account for the complex ways in which algorithmic indirect differential treatment can cause harm.

10. Concluding remarks

Over the course of this article, I have attempted to first precisely define a particular issue that arises in the use of automated decision-systems (ADS), to wit that ADS can indirectly discriminate. I have sketched how the issue emerges, and explored three parity conditions found in the literature on algorithmic fairness. I then used levelling down-type objections to show that we have no moral reason to satisfy any of the parity conditions. Finally, I have argued that that we can explain the moral badness of this form of discrimination by reference to the fact that it causes harm, once we take proper account of the way harm varies across decision scenarios and the persons and groups subject to the ADS. This harm-based account, I have suggested, is at once theoretically plausible and allows us to explain the intuitive differences we respond to in important cases.

Let me make two brief observations on some limitations of the argument. First, while I hope to have shown that the badness of algorithmic indirect discrimination can be explained with reference to the harm that discrimination causes, I have not shown that harm is the only thing that can make indirect algorithmic differential treatment morally bad. The harm-based account is only one general account of what makes

discrimination morally bad.[21] My current view, which I cannot develop here, is that the most prominent alternatives to the harm-based account of discrimination are vulnerable to powerful general objections, and will in some cases fit poorly with the context of ADS.[22] However, those attracted to alternative accounts should be able to accept the present argument while maintaining that algorithmic indirect discrimination is in at least some situations morally (even more) bad for other reasons.

Second, what is the feasibility of employing the analysis I have developed to avoid indirect algorithmic discrimination? If the analysis stands, then training an ADS to respect the ethical demands of appropriate concern for the wellbeing of persons is a more complex task than has been widely recognised. Training must be sensitive to the way the value of classifications can differ across the true/false and positive/negative divisions, and adjust the objective of avoiding the relevant classifications accordingly. Furthermore, developers must consider how the value of errors can vary from one person to another, particularly across the groups conventionally at the heart of our concern for discrimination. When such differences exist, training of the ADS must take into account these differences in order to minimise harm. This will often require "treatment disparity" of some form, i.e. direct differential treatment. (Lipton, Chouldechova, & McAuley, 2018) Finally, the ADS must take into account group vulnerability and its effects, such as stigmatic harm. When more equal treatment has the effect of decreasing the harms resulting from group vulnerability, then developers have reason to promote equality to the point where further decreases in these harms are exceeded e.g. by increased harms resulting from loss of accuracy.

[21] The most prominent alternative accounts in the literature explain the badness of discrimination with reference to disrespect or inequality. Proponents of respect-based accounts argue that discrimination involves a failure to treat persons in the light of reasons grounded in their moral worth (Alexander, 1992; Eidelson, 2015; Glasgow, 2009; Moreau, 2020; Slavny & Parr, 2015), or that it involves treating persons in a way that expresses a demeaning underestimation of their worth (Hellman, 2008). Equality-based accounts hold that discrimination involves a decrease in the wellbeing or life opportunities of persons who are already disadvantaged through no fault of their own. (Knight, 2017; Segall, 2012)

[22] For example, accounts that rely on the discriminator's mental state are likely to fit poorly with ADS that does not have mental states. For critical discussion of disrespect-based accounts, see (Arneson, 2017; Beeghly, 2017; Lippert-Rasmussen, 2006, 2013, 2018a). For critical discussion of the expressive disrespect account, see (Arneson, 2013, pp. 91-94; Eidelson, 2015, pp. 84-90; Lippert-Rasmussen, 2013). For critical discussion of equality-based accounts, see (Lippert-Rasmussen, 2013)

Is it realistic that developers can train ADS that accurately take these factors into account to minimise harm? In many situations, probably not. Developers will typically work under constraints of imperfect information and limited resources. As such, developers will often by necessity train ADSs that cause at least some harm through algorithmic indirect discrimination.[23]

The inevitability of some harms should not induce us to defeatism. If developers employ the understanding developed in the preceding train the ADS so as to minimise the harm done to the best of their ability, then not only will they have done all we can reasonably demand, they will also potentially have protected many vulnerable persons from algorithmic harm. In an imperfect world, that would be no mean ethical achievement.

[23] Recent research on how to develop ADS under constraints sensitive to benefits, welfare and harm, includes (M. Altman et al., 2018; Corbett-Davies et al., 2017; Corbett-Davies & Goel, 2018; Heidari et al., 2019; Speicher et al., 2018)

References

AccessNow. (2018). *Human Rights in the Age of Artificial Intelligence*. Retrieved from https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf

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

Altman, A. (2020). Discrimination. In E. N. Zalta (Ed.), *Stanford Encyclopedia of Philosophy*.

Altman, M., Wood, A., & Vayena, E. (2018). A Harm-Reduction Framework for Algorithmic Fairness. *IEEE Security & Privacy, 16*(3), 34-45. doi:10.1109/MSP.2018.2701149

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* (pp. 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* (pp. 151-163). London: Routledge.

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and machine-learning*. Retrieved from https://fairmlbook.org/

Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. *California Law Review, 104*(3), 671-732. doi:10.2139/ssrn.2477899

Beeghly, E. (2017). Discrimination & Disrespect. In K. Lippert-Rasmussen (Ed.), *Routledge Handbook to the Ethics of Discrimination* (pp. 83 - 96): Routledge.

Benner, A. D., Wang, Y., Shen, Y., Boyle, A. E., Polk, R., & Cheng, Y.-P. (2018). Racial/ethnic discrimination and well-being during adolescence: A meta-analytic review. *American Psychologist, 73*(7), 855-883. doi:https://doi.org/10.1037/amp0000204

Berger, M., & Sarnyai, Z. (2015). "More than skin deep": stress neurobiology and mental health consequences of racial discrimination. *Stress, 18*(1), 1-10. doi:10.3109/10253890.2014.989204

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2018). Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research, Online first*. doi:10.1177/0049124118782533

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

Broome, J. (1990). Fairness. *Proceedings of the Aristotelian Society, 91*, 87-101.

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. doi:10.1017/S1744552319000077

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data, 5*(2), 153-163. Retrieved from https://ui.adsabs.harvard.edu/\#abs/2016arXiv161007524C

Chouldechova, A., & Roth, A. (2018). The Frontiers of Fairness in Machine Learning. *arXiv e-prints*. Retrieved from https://ui.adsabs.harvard.edu/\#abs/2018arXiv181008810C

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

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017, January 01, 2017). *Algorithmic decision making and the cost of fairness.* Paper presented at the KDD '17.

Cosette-Lefebvre, H. (2020). Direct and Indirect Discrimination. *Public Affairs Quarterly, 34*(4), 340-367.

Crisp, R. (2011). In Defence of the Priority View: A Response to Otsuka and Voorhoeve. *Utilitas, 23*(1), 105-108. doi:10.1017/S0953820810000488

Daniels, N. (2008). *Just Health: Meeting Health Needs Fairly*: Cambridge University Press.

Dieterich, W., Mendoza, C., & Brennan, T. (2016). *COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity*. Retrieved from

Donohue, M. (2019). A Replacement for Justitia's Scales? Machine Learning's Role in Sentencing. *Harvard Journal of Law and Technology, 32*(2), 657-678.

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

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2011). Fairness Through Awareness. *arXiv:1104.3913 [cs]*. Retrieved from http://arxiv.org/abs/1104.3913

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

Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., & Venkatasubramanian, S. (2017). *Runaway Feedback Loops in Predictive Policing*. Paper presented at the 1st Conference on Fairness, Accountability and Transparency. https://arxiv.org/abs/1706.09847

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

European Group on Ethics in Science and New Technologies. (2018). *Artificial Intelligence, Robotics and 'Autonomous' Systems*. Retrieved from https://ec.europa.eu/research/ege/pdf/ege_ai_statement_2018.pdf

Ferguson, A. G. (2017). *The Rise of Big Data Policing: Surveillance, Race, and the Future of Law Enforcement*: NYU Press.

FRA. (2018). *#BigData: Discrimination in data-supported decision making*. Retrieved from http://fra.europa.eu/sites/default/files/fra_uploads/fra-2018-focus-big-data_en.pdf

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*. Paper presented at the Proceedings of the Conference on Fairness, Accountability, and Transparency, Atlanta, GA, USA. https://doi.org/10.1145/3287560.3287589

Glasgow, J. (2009). Racism as Disrespect. *Ethics, 120*, 64-93.

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, Online first*.

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. (2020). Measuring Algorithmic Fairness. *Virginia Law Review, 106*(4), 811-866.

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

High-Level Expert Group on Artificial Intelligence. (2019). *Ethics Guidelines for Trustworthy AI*. Retrieved from https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=58477

Holtug, N. (2010). *Persons, Interests, and Justice*. Oxford: Oxford University Press.

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

Jaume-Palasí, L., & Spielkamp, M. (2017). *Ethics and algorithmic processes for decision making and decision support*. Retrieved from https://algorithmwatch.org/wp-content/uploads/2017/06/Ethik_und_algo_EN_final.pdf

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

Khaitan, T. (2017). Indirect Discrimination. In K. Lippert-Rasmussen (Ed.), *Routledge Handbook of the Ethics of Discrimination* (pp. 30-41): Routledge.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2017). Human Decisions and Machine Predictions. *NBER Working paper series*. Retrieved from http://www.nber.org/papers/w23180

Kleinberg, J., Ludwig, J., Mullainathan, S., & Sunstein, C. R. (2019). Discrimination in the Age of Algorithms. *arXiv e-prints*. Retrieved from https://ui.adsabs.harvard.edu/\#abs/2019arXiv190203731K

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent Trade-Offs in the Fair Determination of Risk Scores. *arXiv e-prints*. Retrieved from https://ui.adsabs.harvard.edu/\#abs/2016arXiv160905807K

Knight, C. (2017). Discrimination and Equality of Opportunity. In K. Lippert-Rasmussen (Ed.), *Routledge Handbook of the Ethics of Discrimination* (pp. 140-150). London: Routledge.

Krieger, N. (2014). Discrimination and Health Inequities. *International Journal of Health Services, 44*(4), 643-710. doi:10.2190/HS.44.4.b

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

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

Lippert-Rasmussen, K. (2007). Private Discrimination: A Prioritarian Desert-Accommodating Account. *San Diego Law Review, 43*, 817-856.

Lippert-Rasmussen, K. (2008). Discrimination and the Aim of Proportional Representation. *Politics, Philosophy & Economics, 7*, 159-182.

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

Lippert-Rasmussen, K. (2015). *Luck Egalitarianism*: Bloomsbury Publishing.

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

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

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

Lipton, Z. C., Chouldechova, A., & McAuley, J. (2018). *Does mitigating ML's impact disparity require treatment disparity?* Paper presented at the 32nd Conference on Neural Information Processing Systems.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2018). Prediction-Based Decisions and Fairness: A Catalogue of Choices, Assumptions, and Definitions. arXiv:1811.07867. Retrieved from https://ui.adsabs.harvard.edu/abs/2018arXiv181107867M

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

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society, 3*(2). doi:10.1177/2053951716679679

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

MSI-AUT. (2018). *A study of the implications of advanced digital technologies (including AI systems) for the concept of responsibility within a human rights framework*. Retrieved from https://rm.coe.int/draft-study-of-the-implications-of-advanced-digital-technologies-inclu/16808ef255

MSI-NET. (2017). *Algorithms and Human Rights - Study on the human rights dimensions of automated data processing techniques and possible regulatory implications*. Retrieved from https://rm.coe.int/study-hr-dimension-of-automated-data-processing-incl-algorithms/168075b94a

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

Otsuka, M., & Voorhoeve, A. (2009). Why It Matters That Some Are Worse off than Others: An Argument against the Priority View. *Philosophy & Public Affairs, 37*(2), 171-199. Retrieved from http://www.jstor.org.ep.fjernadgang.kb.dk/stable/40212842

Panel for the Future of Science and Technology. (2019). *Understanding algorithmic decision-making: Opportunities and challenges*. Retrieved from

Parfit, D. (2002). Equality or Priority. In M. Clayton & A. Williams (Eds.), *The Ideal of Equality* (pp. 81-125). Basingstoke: Palgrave Macmillan.

Parfit, D. (2012). Another Defence of the Priority View. *Utilitas, 24*(3), 399-440. doi:10.1017/S095382081200009X

Perry, W. L. (2013). *Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations*: RAND Corporation.

Rainie, L., & Anderson, J. (2017). *Code-Dependent: Pros and Cons of the Algorithm Age*. Retrieved from http://www.elon.edu/docs/e-web/imagining/surveys/2016_survey/Pew%20and%20Elon%20University%20Algorithms%20Report%20Future%20of%20Internet%202.8.17.pdf

Rawls, J. (1999). *A Theory of Justice*. Oxford: Oxford University Press.

Reisman, D., Schultz, J., Crawford, K., & Whittaker, M. (2018). *Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability*. Retrieved from https://ainowinstitute.org/aiareport2018.pdf

Roth, A. (2016). Trial by Machine. *Georgetown Law Journal, 104*(5), 1245-1306.

Schmitt, M. T., Branscombe, N. R., Postmes, T., & Garcia, A. (2014). The consequences of perceived discrimination for psychological well-being: A meta-analytic review. *Psychol Bull, 140*(4), 921-948. doi:https://doi.org/10.1037/a0035754

Segall, S. (2012). What's so bad about Discrimination? *Utilitas, 24*(1), 82-100.

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

Speicher, T., Heidari, H., Grgic-Hlaca, N., Gummadi, K. P., Singla, A., Weller, A., & Zafar, M. B. (2018). *A Unified Approach to Quantifying Algorithmic Unfairness: Measuring Individual &Group Unfairness via Inequality Indices*. Paper presented at the Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, United Kingdom. https://doi.org/10.1145/3219819.3220046

Temkin, L. S. (2002). Equality, Priority, and the Levelling Down Objection. In M. Clayton & A. Williams (Eds.), *The Ideal of Equality* (pp. 126-161). Basingstoke: Palgrave Macmillan.

Voorhoeve, A., & Fleurbaey, M. (2012). Egalitarianism and the Separateness of Persons. *Utilitas, 24*(3), 381-398. doi:10.1017/S0953820812000040

Williams, D. R., Lawrence, J. A., Davis, B. A., & Vu, C. (2019). Understanding how discrimination can affect health. *Health Services Research, 54*(S2), 1374-1388. doi:https://doi.org/10.1111/1475-6773.13222

Zarsky, T. (2016). The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making. *Science, Technology & Human Values, 41*(1), 118-132. doi:10.1177/0162243915605575

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

Zuiderveen Borgesius, F. (2018). *Discrimination, artificial intelligence, and algorithmic decision-making*. Retrieved from https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73