Two Kinds of Discrimination in AI-Based Penal Decision-Making

Dietmar Hübner
Institute of Philosophy, Leibniz University Hannover
Im Moore 21, D-30167 Hannover, Germany
dietmar.huebner@philos.uni-hannover.de

ABSTRACT

The famous COMPAS case has demonstrated the difficulties in identifying and combatting bias and discrimination in AI-based penal decision-making. In this paper, I distinguish two kinds of discrimination that need to be addressed in this context. The first is related to the well-known problem of inevitable trade-offs between incompatible accounts of *statistical fairness*, while the second refers to the specific standards of *discursive fairness* that apply when basing human decisions on empirical evidence. I will sketch the essential requirements of non-discriminatory action within the penal sector for each dimension. Concerning the former, we must consider the relevant *causes* of perceived correlations between race and recidivism in order to assess the moral adequacy of alternative standards of statistical fairness, whereas regarding the latter, we must analyze the specific *reasons* owed in penal trials in order to establish what types of information must be provided when justifying court decisions through AI evidence. Both positions are defended against alternative views which try to circumvent discussions of statistical fairness or which tend to downplay the demands of discursive fairness, respectively.

Keywords

AI-based decision-making, crime prediction, bias, discrimination, justice, fairness, statistical fairness, discursive fairness, COMPAS.

1. INTRODUCTION

This paper addresses two types of potential discrimination in algorithm-based decision-making. Both problems are well-known and have achieved widespread attention within the general public and the academic disciplines, though it may be surprising that I discuss both of them under the title "discrimination". However, I hope the following discussion will elucidate why this overarching perspective is conceptually justified and ethically insightful. In particular, it may clarify central questions within both dimensions and reveal important connections between the two issues.

Concerns of possible bias and discrimination in computer algorithms pertain to a multitude of areas, ranging from everyday applications such as image recognition software, search engines or chat bots to specialized systems used in university admission procedures, hiring decisions or loan granting. Notwithstanding this ubiquity of the topic, it is plausible to assume that normative analyses of and technical solutions to bias and discrimination in AI-based decision-making must ultimately be tailored to the concrete fields of social interaction in which these applications take place. In this paper, I will focus on the forensic sector, more specifically on AI-based crime prediction in penal court decisions. My primary example of use will be the famous COMPAS case.

COMPAS ("Correctional Offender Management Profiling for Alternative Sanctions") is a software package for crime prediction which was originally developed and marketed by the private firm Northpointe, meanwhile succeeded by Equivant (www.equivant.com). COMPAS aims to assess a defendant's probability of reoffending, thus supporting judges in their decisions concerning whether a culprit should be detained before trial or might be released on bail, whether she should go to prison or might be eligible for probation, or whether she should stay in jail or might be a candidate for parole [13, 31]. COMPAS is based on a questionnaire, collecting data such as the current charges against the defendant and her criminal history, but also socio-economic factors including education levels, employment status, family background, social environment etc. Drawing from 137 items, COMPAS generates a score predicting the risk of reoffending, ranging from 1 to 10 [9].

In 2016, a public controversy arose when Angwin and colleagues claimed that COMPAS risk scores were discriminating against black persons, pointing to apparent problems of *statistical fairness* in the algorithm's predictions [3, 28]. Beforehand, several authors had already expressed *procedural worries* about using AI evidence in court hearings [8, 10]. In the following sections, I will turn to both aspects respectively, highlighting the basic tenets of how each notion of fairness is to be assessed. Finally, I will connect these analyses into an integrated view.

2. STATISTICAL FAIRNESS

2.1 Numerical Features of COMPAS

Angwin et al.'s statistical concerns about COMPAS can be most readily retraced by arranging retrospective data on algorithmic predictions and real outcomes in an "error matrix" (or "confusion matrix"). In such a matrix, rows contain the numbers of persons predicted to exhibit a certain trait (here: predicted reoffending or predicted not reoffending), while columns display the numbers of persons indeed falling into the corresponding groups (here: in fact reoffending or in fact not reoffending). The four resulting fields of the matrix contain true positives (TP: predicted reoffending and in fact reoffending), false positives (FP: predicted reoffending, but in fact not reoffending), true negatives (TN: predicted not reoffending and in fact not reoffending), and false negatives (FN: predicted not reoffending, but in fact reoffending). Based on these numbers, several parameters can be calculated in order to evaluate the statistical performance of the algorithm.

Issues of statistical fairness can be addressed by calculating these parameters separately for different groups. In particular, discrepancies in parameters between groups distinguished by salient features such as race or gender may be taken as indications of potential discrimination. The following error matrix for COMPAS (Table 1), compiled from Larson et al., shows algorithmic predictions

and real outcomes in Broward County, Florida (2013/14), contrasting data on white (w) and black (b) defendants [28]. By analyzing these basic data, especially by comparing statistical parameters for whites and blacks, one can determine whether COMPAS satisfies different fairness conceptions [5, 6, 19, 27, 30, 34].

	In fact reoffending	In fact not reoffending
Predicted reoffending	TP w = 505; b = 1,369	FP w = 349; b = 805
Predicted not reoffending	FN w = 461; b = 532	TN w = 1,139; b = 990

Table 1: Error matrix for COMPAS, numbers from [28][1]

First, it must be stressed that COMPAS does not make *explicit use* of a *race variable* in order to generate its predictions. Race is not among the 137 items on the questionnaire, and nothing suggests that COMPAS reconstructs its value from proxy variables and then utilizes it as an additional input to its calculations. So COMPAS does comply with the standard of "fairness through unawareness", not referring to a variable that would carry the label "protected" in fields pertaining to possible racial discrimination. However, it is widely agreed that this is a minimum requirement which, in general, does not exhaust all statistical fairness issues.

Second, *predicted rates (PRs)*, i.e. relative numbers of persons predicted to reoffend, can be calculated (PR = {TP+FP}/total). Here we find a stark difference between both groups ($PR_w = 35\%$, $PR_b = 59\%$), demonstrating that COMPAS does not correspond to the standard of "statistical parity" ("demographic parity", "equal acceptance rate"). However, one might consider it adequate to compare the PRs to the *true rates p* ("base rates", "prevalences"), i.e. the relative numbers of individuals who do in fact reoffend (p = {TP+FN}/total). Although not in perfect agreement, these display at least a similar tendency ($p_w = 39\%$, $p_b = 51\%$). Against this background, it may appear incongruous to complain about unequal predicted rates PR. Rather, it might be suggested, COMPAS simply tracks social reality, as displayed in the true rates p. I will comment on this issue in the following sections, particularly on the problems of calling the prevalences p "true rates". But for the time being, it should be noted that also Angwin et al., in their critique of COMPAS, do not focus on its lack of statistical parity.

Instead, what Angwin et al. are predominantly concerned about is the *false positive rates (FPRs)*, i.e. the relative numbers of persons who, although they ultimately do not reoffend (i.e. being either a false positive or a true negative), have erroneously been predicted to reoffend (ending up as a false positive) (FPR = FP/{FP+TN}). This indicator is much higher for blacks than for whites ($FPR_w = 23\%$, $FPR_b = 45\%$), i.e. COMPAS does not fulfil "predictive equality". Maybe not surprisingly, the reverse is true for the *false negative rates (FNRs)*, i.e. the relative numbers of individuals who, although they ultimately do reoffend (i.e. being either a true positive or a false negative), have erroneously been predicted not to reoffend (ending up as a false negative) (FNR = FN/{TP+FN}). This figure is much higher for whites than for blacks ($FNR_w = 48\%$, $FNR_b = 28\%$), i.e. COMPAS does not satisfy "equal opportunity". So in both regards, we do not have "error rate balance", and taken together we do not have "equalized odds", which would require both error rates to be equal. Put simply, COMPAS seems to be too strict for blacks and too lax for whites.

However, Northpointe replied to these concerns by stating that this difference should not be misread as racial bias against black defendants [11]. In particular, they argued that the appropriate metric for judging fairness is rather the *positive predictive values (PPVs)*. This parameter measures the relative numbers of persons who, after having been predicted to reoffend (i.e. being either a true positive or a false positive), do in fact reoffend (ending up as a true positive) (PPV = TP/{TP+FP}). This value, although not fully identical, is in reasonable agreement for both groups ($PPV_w = 59\%$, $PPV_b = 63\%$). So COMPAS does establish approximate "predictive parity" (essentially equivalent to "calibration"). In this respect, COMPAS does not seem to discriminate against black people.

It appears like a natural requirement that all the above parameters, error rates as well as predictive values, be roughly equal for whites and blacks in order to avoid potential discrimination. But unfortunately, this comprehensible demand, except for degenerate cases like zero errors, is mathematically impossible to meet, where the prevalences p differ for both groups. There are several "impossibility theorems" demonstrating this unfavorable constellation [16, 25]. Maybe the most easily accessible proof is by Chouldechova [7]. She bases her argument on the equation FPR = [p/{1–p}] · [{1–PPV}/PPV] · [1–FNR]. From this formula, it can easily be seen that, if there is a difference in prevalence p for the two groups, the groups must also differ in at least one of the three quality parameters, FPR, FNR or PPV, unless these have values zero or one.[2]

2.2 A Stalemate

Against this background, some scholars have started to turn away from the debate on statistical fairness, preferring other approaches to issues of algorithmic discrimination [6, 18, 19, 24]. This reaction is comprehensible, and appears to be backed by several considerations.

First, the impossibility theorems mentioned above demonstrate that we cannot have all that we might want in terms of statistical fairness. And facing this irresolvable conflict of alternative conceptions, it is not obvious which fairness measure to prefer.

[1] Larson et al. only analyzed pretrial-detainment decisions, not probation or parole decisions, as COMPAS was predominantly used for the former in Broward County. They classified as "predicted reoffending" individuals receiving a risk score of 8–10, and as "predicted not reoffending" those with a risk score of 1–4, corresponding to Northpointe's classification of these individuals as "high risk" or "low risk", respectively. They defined "in fact reoffending" or "in fact not reoffending" with regard to whether the same person was again arrested within two years after her scoring, as COMPAS itself is supposed to predict a new offence within two years. I will comment on the problem of identifying reoffending with rearrests in due course.

[2] COMPAS, in fact, has significant differences in two parameters. Both FPRs and FNRs considerably deviate for whites and blacks. Theoretically, an algorithm could achieve equality of two parameters between the groups. But at least one of the three needs to compensate for the given difference in prevalence p.

To be sure, there is some reason to share Angwin et al.'s view that the differing FPRs for whites and blacks are especially disturbing. COMPAS is applied within the penal system, where false positives appear to be particularly troubling. It may be tempting to back this normative intuition with reference to the basic standard *in dubio pro reo*. However, this classical legal tenet concerns the *ascription of past offences* to a defendant: it states that, if you are not reasonably sure that some person has committed a crime, she should rather not be prosecuted. COMPAS, by contrast, is applied in decisions concerning bail, probation or parole, where the *prediction of future offences* of the defendant is at stake: in these contexts, we are fairly certain that a person has committed some offence, but we consider waiving incarceration, given an optimistic prediction that she will not perpetrate again. Consequently, *in dubio pro reo* does not properly apply here. In particular, false positives in COMPAS do not, as is sometimes suggested, amount to "incarcerating innocent people". If that was the case, even an FPR of 23% for whites would be outrageously high, not just an FPR of 45% for blacks. Yet, it may be reasonable to hold that, although the *in dubio* principle itself is not to the point, some more basic imperative standing behind it will still apply, namely the idea that it is legally paramount to avoid unnecessary punishment. Even when defendants are highly suspect or actually convicted of some past offence, imprisonment without demonstrated need ought to be avoided where possible, given its devastating impact on individuals and their families. Consequently, once we agree that the past offence in question is of a minor kind or has been atoned for to a sufficient degree, so that good confidence in the future compliance of a defendant would justify her release, failure to grant bail, probation or parole would constitute a major wrong within a liberal state, ultimately conflicting with the rule of law. Following this line of thought, focusing on FPRs, rather than on FNRs or PPVs, would appear paramount to penal justice. Admittedly, though, it may be less obvious why the FPRs need to be equal for different groups. We should possibly *minimize* them, but it is not clear yet why we should *equalize* them.

Similar remarks hold with regard to the FNRs. Contrary to the arguments sketched above, one may insist that, in discussions of bail, probation or parole, false negatives should constitute our primary focus. In these contexts, judges are presented with highly suspect or actually convicted individuals whose incarceration would be basically justified. Under these circumstances, decisions to waive imprisonment must, first and foremost, avoid possible dangers to the general public due to potentially non-compliant, dangerous, recidivating individuals. This is why predictions of future offences are involved in these decisions. The defendants did commit an offence, or are highly suspect of having done so, and the question of whether imprisonment could be waived must focus more on the danger of future recidivism in case of false negatives than on the danger of unnecessary imprisonment in the form of false positives. Note that this argument would not, as might first appear, establish that there was no race-related problem in COMPAS: to be sure, it would shift the focus away from the disproportionate numbers of false positives in black defendants that Angwin et al. concentrate on. But instead, it would have to turn to the enlarged numbers of false negatives in white defendants: stressing the need to protect the public, of all things an FNR of 48% for whites would seem to be unbearably high, not so much an FNR of 28% for blacks. Again, however, this line of reasoning may not really bear on issues of statistical fairness. It would probably require the *minimization* of FNRs, but it may not straightforwardly suggest the importance of their *equalization*.

It is also understandable that Northpointe underlined the importance of PPVs. Given that the algorithm predicted that a person would reoffend, the PPV indicates the probability that the person will indeed do so. So in a way, the PPV announces the reliability of the algorithmic prediction, the quality of the provided service. Thus, it is not surprising that computer scientists are inclined to focus on this parameter, and that common processes of algorithm optimization tend to increase its value. In addition, the information conveyed by the PPV seems to be in better correspondence to the epistemic situation of a decision-maker than the FPR or FNR. She is not presented with a not reoffending or reoffending individual and has to make up her mind whether the algorithm might misclassify that person (FPR or FNR), but she is presented with an algorithmic prediction and has to make up her mind whether this assessment will turn out to be true (PPV). Finally, it makes sense to assume that the PPV should not only be *maximized*, but also *equalized* across groups. For, if this is the case, the decision-maker (i.e. the judge) may restrict her considerations to the prediction that she is given (i.e. the risk score), without having to pay additional attention to the defendant's group affiliation when interpreting this information. If the PPV is equal for whites and blacks, a given risk score has the same meaning for both groups. The prediction has a consistent reliability, no matter whether the person concerned is white or black.

In short, there seems to be a real stalemate between these different fairness measures. And it may appear hopeless to find a decisive argument in favor of one of them.

Second, focusing on measures of statistical fairness runs the danger of absurd solutions, ending up with an AI that simply rearranges numbers in the error matrix in the way desired, but without any substantial sense [12, 18]. For instance, an algorithm could achieve statistical parity by predicting proportionate fractions from two groups to reoffend while selecting the individuals from both groups randomly, or it could equalize false positive rates by attributing high risk scores to actually harmless individuals.

Third, largely analogous debates on different statistical fairness standards and their mutual mathematical incompatibility took place back in the 1960s and 1970s, in discussions on potential bias and discrimination in assessment tests [21]. These discussions produced no decisive results, undermining hopes that we might do now better with the parallel problems in algorithmic predictions.

Given these findings, it is not surprising that some people have become weary of discussions on statistical fairness. At the same time, something important still seems to show up in the numbers which is worth addressing. In the following sections, I will make some remarks on these issues and suggest how they might be tackled. In particular, I contend that there is no general solution stating which fairness measures should dominate in any AI-based decision-scenario, be it university admissions or loan granting, but that we need to turn to a concrete scenario, like crime prediction in the forensic sector, in order to approach these problems.

2.3 The Core Question

To adequately grasp the issue of statistical fairness in AI-based penal decision-making, one core question needs to be addressed: *What is the cause of the correlation between race and recidivism that we appear to observe both in empirical data and in AI predictions?* It is only answers to this core question, I propose, that can guide us in balancing various statistical fairness demands. Two main answers to this question seem to suggest themselves.

(1) The first answer would be: "A major cause of the correlation is the *past treatment* of black people in the US. In the US history, we witness an extensive thread of *massive discrimination* against black persons, including slavery, political exclusion, segregation, and social marginalization. This practice has clearly led to a significant *socio-economic deprivation* of the black population. And higher crime incidence, or enlarged recidivism rates, as they show up both in empirical data and in AI predictions, must be regarded, to a large extent, as another *downstream effect* of this targeted maltreatment."

This attitude would assume that there is a *true correlation* between race and recidivism in social reality, i.e. that there is in fact a higher rate of reoffending in black defendants. But it would underline that this correlation is an obvious effect of *past wrongs* done to that population. In this paper, I will not try to enter into a political debate whether this perspective is adequate. Recent reports on systemic racism in the US police may suggest that higher crime rates in the black population are, to a considerable degree, a myth [4].[3]

For my current purpose, however, I want to explore what reactions this diagnosis would entail. And I think what suggests itself would be the idea that some kind of "affirmative action" might be applicable here [17]. Affirmative action comprises political measures meant to counter the disproportionate prevalence of salient groups in certain areas of public life. Such programs have been mainly justified in two different ways.

Firstly, and predominantly, affirmative action is grounded on the aim to promote diversity, plurality, integration or participation, e.g. in classrooms, universities, workplaces and offices. The driving idea behind this conception is recognition of the fact that these social units themselves *benefit* from the presence of different experiences and world views, and that society at large *needs* e.g. black attorneys or female managers in order to retain social cohesion and provide role models. However, it seems dubious how this line of justification might apply to the case at hand. It would appear strange to argue that we need racial diversity or racial integration in person imprisoned or in persons being released.

Secondly, however, affirmative action can be justified through the definite purpose to counter social correlations based on acknowledged wrongs. The main idea behind this conception is that we should ignore or override certain criteria that we usually apply in our assessments if it should turn out that they are tainted by past discrimination, in order to prevent these *past wrongs* from further infecting our *current decisions*. For instance, when we find that test results correlate with race or gender, and when we know that these correlations obtain because blacks or women have been subjected to preceding discrimination in their development and education, we should suppress or overrule these indicators, at least to some extent, and accept those applicants, in spite of their poorer performances. We should compensate them, not in the cheap sense of giving them some arbitrary advantages in order to balance their former harms, but in the conscientious sense of not letting their past disadvantages determine their future fates.

Following this line of thought, affirmative action advocates the targeted departure from common decision criteria in order to prevent past discrimination from influencing people's future lives. Within the context of COMPAS, this would mean that predicted rates for recidivism, when underlying judgements on bail, probation or parole, should be taken at values deviating from the true rates, correcting them for their problematic background in past injustice. More precisely, the *predicted rates* should be taken as equal, or at least more equal. Consequently, *statistical parity* would be the fairness measure to adhere to, at least to some extent [6, 12, 14, 15].

There may be some debate concerning whether this perspective is persuasive. For instance, the compensatory approach to affirmative action, as opposed to the diversity logic, presupposes that the concrete individual, and not just her social group, has been personally affected by the past wrongs in order to justify her favorable treatment, which might be hard to argue for in a given case of penal justice [17]. Moreover, it may be doubted that abstaining from punishment can really count as compensating for disadvantages, comparable to offering someone a university place considering her deficient education. In addition, our reason for ignoring poor test results may ultimately be backed by our confidence that the person thus favored might eventually succeed at our university, hoping that her hitherto underdeveloped talents will be awakened through high quality teaching, whereas in ignoring high risk scores we would have to acknowledge that we do in fact underrate her probability of reoffending, as her personality structure is likely to fail again in her unimproved circumstances.

Notwithstanding these caveats, affirmative action is basically applicable to any social system. And the demand to eradicate the social influence of past wrongs has at least some argumentative weight in the penal context. At any rate, it should be noted that the concept is not meant to apply to extremely dangerous criminals expected to commit further violent felonies. Its use is restricted to persons who, given their past and present record, are realistically eligible for bail, probation or parole.

(2) A different answer would be: "A major cause of the correlation is the *current treatment* of black people in the US. In the US criminal system, we find a systematic policy of *racial targeting* of black people, consisting in more intensive surveillance, more frequent arrests, and more severe sentences. This skewed practice leads to *false data* in the training sets from which COMPAS has learned, and these exaggerated trends are now being reproduced in the algorithm's predictions. In particular, the higher 'prevalence', the differing 'base rate' or disproportionate 'true rate' that seems to show up in retrospective assessments, is actually, to a large extent, a *social artefact* and not 'true' at all."

One major problem that this position will emphasize is the fact that, while COMPAS is generally supposed to predict future *offences*, as only the probability of impending offences can have any legitimate impact on court decisions concerning bail, probation or parole, COMPAS is actually designed to predict future *arrests*, as a closer reading of the official Practitioner's Guides reveals,
simply because it has been mainly trained on data sets of past arrests [9, 13, 31]. This implicit equation of (re-)offending with (re-)arrests in the application of COMPAS is plainly *wrong*, as not every arrest is based on a verified offence, and it is clearly *biased*, as in the US blacks are much more likely than whites to be subject to unfounded arrests without having committed an actual offence [4]. Using COMPAS, however, will feed this bias back into the system and perpetuate it. Based on false data (disproportionate arrest rates), it will make distorted predictions (concerning future offences), thus producing enlarged imprisonment rates, thus suggesting exaggerated crime rates, thus encouraging more racial

[3] I will come back to this skepticism below. Essentially, it converges with the alternative answer to the core question.

targeting, thus generating more false data, thus making more distorted predictions, and so on [22, 27].[4]

However, if it is false data that underlie our decisions, differing false positive rates are particularly hard to accept. In any case, unnecessary punishment is a major problem for penal justice, but if it is based on false data, it becomes clearly untenable.

In this light, Angwin et al.'s focus on the false positive rates is most comprehensible. As stated above, we may generally debate whether false positives, false negatives or positive predictive values are of paramount importance in penal justice. But when we learn that *unnecessary imprisonments* stem from constant misinformation, *false positives* must become our major concern.

In addition, against this background it makes sense not just to demand the minimization of false positive rates, but also their equalization across groups. When data are distorted to the detriment of one group, resulting differing error rates become a real issue. When *deviating miscarriages* are based on fake differences, we must avoid *differing mistakes* in harming people.

Correspondingly, this second answer to the core question suggests that our major concern should in fact be to equalize *false positive rates*. In technical terms, our algorithm should strive to satisfy *predictive equality*, rather than one of the other fairness measures [20, 35].

There is a little problem with this conclusion, as it apparently presupposes the FPRs to be objectively true when calling for their equalization. If these numbers are themselves infected by false data, equal or minimal or even zero FPRs will be no real comfort as they will still perpetuate the current discrimination in the system [6]. And in fact, we must suspect that FPRs, as reported by Angwin et al., are still distorted, because they follow COMPAS in counting rearrests as reoffences. So not only the "true rate" is not "true", as the second answer stresses, but the FPRs are not true either, although the second argument seeks to equalize them.

However, this inconsistency does not undermine the argument in a fatal way. Admittedly, the call for equal FPRs should ultimately not apply to Angwin et al.'s own figures, but to ideal numbers, counting as true positives or false negatives not simply all rearrested persons, but only individuals who do indeed reoffend. But this caveat does not contradict the basic idea that false positives must be the major concern against the background of biased training data. And if the FPR in blacks is too high even for the distorted numbers, at least that obvious mistake should be reduced, all the more as we have to suspect that their true FPR is bigger still, containing all the rearrested persons who did not reoffend.

2.4 Division of Labor

I will not try to explore how adequate the two answers to the core question are, or decide which argument is more convincing. It seems reasonable to assume that both asserted ways of influence contribute to the situation, and that both suggested remedies can be supported: There may be some true correlation between race and recidivism, based on past discrimination, which can encourage the affirmative action logic and thus make us want to move towards more equal predicted rates. There may also be false data underlying the algorithmic predictions, based on racial targeting, which should make unnecessary imprisonment our major concern and hence call for more equal false positive rates.

Even if both lines of reasoning apply, though, it is helpful to highlight their divergent focuses and disentangle their logical structures. Not least, this differentiation may be important in deciding which corrections should be performed by which player, suggesting a division of labor: Carrying out compensatory adjustments to predicted rates in the spirit of affirmative action might ultimately be the business of human users at the end of the decision-making process, and so best be realized by the judges: this conforms to widespread intuitions that it is up to society, and not to the algorithms, to take charge of correcting the long-term effects of discriminatory practices that shape our communities [19]. Balancing out error rates due to false data, by contrast, should rather be regarded as part of the algorithmic service provided, and so be taken care of by the programmers: correcting for problematic input should take place before the predictive output is presented.

At the same time, this ideal disentanglement may have its realistic limits. We must be prepared to meet deeper intertwinements between the two lines of argument, at all levels from diagnoses to principles and remedies.

Factual assessment. In an indirect sense, past discrimination may as well contribute to false data: ultimately, it is these historical practices that have brought about present stereotyping, prejudice and harshness on the side of police agencies. Conversely, current racial targeting may to some extent contribute to true correlations [19]: in fact, by generating opposition, resignation or role-acceptance within the black population it may reinforce problematic behavioral patterns.

Normative demands. In a certain sense, statistical parity, i.e. the aim of having more equal predicted rates, may also be seen as an approximate correction of skewing effects due to racial targeting [15]: in any case, deviation from observed correlations due to affirmative action is easier to accept when it is clear that these allegedly "true rates" are not "true" at all. Conversely, predictive equality, i.e. the aim of having more equal false positive rates, may be regarded a minimum requirement of not adding further wrongs to past discrimination: after all, the unnecessary punishment of black people appears like an unpleasant continuation of malpractices such as slavery, exclusion, segregation or marginalization.

Technical implementation. When applying common techniques to "debias" algorithms against the background of differing base rates, we might expect to produce both corrections simultaneously. Especially when not performing some selective rearrangement of entries in the error matrix, but following a more reasoned approach (e.g., by looking for variables strongly correlating with race and eliminating these variables from the training set), the ultimate effect will be to assimilate groups, i.e. more equal predicted rates and more equal error rates. This will usually have its costs, because eliminating information from the data will generally impair the accuracy of the algorithm (e.g., the error rates will be more equal, but they will go up). But this is the kind of price you always pay for affirmative action, and if the information is dubious anyway it may not be a high price.

So there is some irony in all of this. Both positions sketched above follow distinct paths of problem assessment and suggested solutions, and also imply a division of labor between judges and programmers in fighting statistical discrimination in AI-based

[4] A defense attorney presented with a high COMPAS risk score of her black client might callously reply, with only a minor admixture of outright sarcasm: "Of course, my client has a high probability of being rearrested – she is black!"

decision-making. But then again, both approaches appear to be deeply merged. Causes and their statistical effects, aims and their moral justification, and techniques and their mathematical impacts seem to be ultimately entangled. We may think that unequal false positive rates are the most urgent problem for a just penal system, particularly as they partly stem from false data due to racial targeting. However, in fighting this phenomenon (preferably at the algorithmic level), we will probably also produce more equal predicted rates, maybe ignoring some real differences between the groups. Now this is what affirmative action (most reasonably applied at the human level) is always about, overriding true correlations that originate in past discrimination. But anyway, past discrimination also contributes to false data and encourages equalizing false positive rates, and likewise, racial targeting enforces true correlations and suggests equalizing predicted rates.

2.5 Alternative Approaches

In Section 2.2, I noted that some scholars have proposed dropping issues of statistical fairness altogether. The alternative approaches to algorithmic discrimination that they instead pursue mostly refer to conceptions of causal reasoning [6, 19, 24], often employing standards of counterfactual fairness [18, 27]. It is beyond the scope of this paper to comment on these strategies in detail. However, it seems likely that their arguments will ultimately depend on considerations largely parallel to those sketched above.

Modern accounts of causal reasoning predominantly refer to Bayesian networks, representing and quantifying deterministic or probabilistic influences between relevant variables of specific systems [32]. In the present context, these conceptions would amount to unearthing the causal paths between (i) the sensitive feature "race", (ii) other attributes collected through the items on the COMPAS questionnaire such as education levels, family background etc., (iii) the risk score arrived at by the algorithm, and (iv) the real outcome of recidivating or not recidivating [6, 19, 24]. On this basis, one could check for discriminatory causal paths within the Bayesian network. In particular, causal paths between the variables "race" and "score" that are deemed illegitimate could be marked out as indicating wrongful discrimination by the algorithm. Obviously, *direct paths* between "race" and "score" would be unacceptable in this sense, as any immediate influence of the protected variable on the algorithmic prediction would amount to straightforward discrimination. However, *indirect paths*, mediated through other variables such as employment status or social environment, could be more controversial, in particular when these mediating variables do impact on recidivating behavior. So which of these paths are to be classified as "discriminatory", and which are to be accepted as legitimate?[5] A rather *extreme position* would regard all indirect paths between "race" and "score" as illegitimate. The effect would be that COMPAS were to be rated as thoroughly discriminatory, because its scores largely depend on such mediating variables. But this attitude seems to amount to an all-too sweeping exculpation of defendants, leaving no notion of personal accountability for any predictive traits that may be statistically correlated with someone's race. A more *moderate position* would consider some of these indirect paths to be tolerable, others less so. The effect of this would be that COMPAS might be in need of some corrections, but not of complete abandonment. Such a view would probably try to distinguish between variables that should be regarded as lying within a defendant's liability and those that she ought not to be accountable for. But how can this line be drawn, without getting deeply entangled in notoriously difficult metaphysical issues of free will and moral responsibility? A promising approach is to concentrate on the distinctly normative dimension of this question, asking which of the correlating variables are linked to race due to plainly unfair practices. More precisely, we need to know whether *past discrimination* has produced stable paths between "race" and "score", and we need to know whether *present discrimination* has produced false values of influences between these variables. This implies that, instead of just putting together a Bayesian network of *causal paths*, we will have to analyze the empirical causes of the *network itself*, in order to assess whether it contains evidence of algorithmic discrimination. But then, causal reasoning on algorithmic discrimination eventually carries us back to exactly those questions concerning affirmative action and unfounded data that our above discussion on statistical fairness has already marked out as essential.

Contemporary arguments on counterfactual fairness may be understood as special variants of causal analysis, arrived at by giving causes a counterfactual interpretation [29]. Within the current debate, these approaches amount to asking whether COMPAS would confer a different risk score to some person if she belonged to a different race [18, 27]. This approach has some intuitive appeal as a guide to discrimination issues. Apparently, an algorithm should be regarded as wrongfully discriminating against a black person if it gave that individual a lower risk score when switching her race variable from "black" to "white". To be sure, the envisaged alteration of her race should not come along with all sorts of *additional changes* within her personality or behavior, as such extra variations might certainly justify corresponding adjustments of algorithmic predictions. Rather, the change must be restricted to a shift in *race alone*, if a difference in score is to be indicative of discrimination. But how precisely is this fictitious state of some person "merely" having a different race to be envisaged?[6] On a rather *naïve interpretation*, it would mean that just the variable "race" changes its value, while all the items contained on the COMPAS questionnaire remain constant [18]. In that case, of course, COMPAS would output the same risk score as before, as it does not use the feature "race" at all, but only the items on the questionnaire. However, to conclude from this fact that there is no problem of discrimination involved appears simplistic, essentially reducing the concept of non-discrimination to plain "fairness through unawareness". On a more *realistic interpretation*, it would mean that the variable "race" changes, and along with it many other items on the COMPAS questionnaire which are socially correlated with race, including education levels, family background etc. [27]. In that case, COMPAS might certainly change its risk score. But it is not obvious that such a change would necessarily indicate discrimination, if we admit that these variables might correlate with criminal behavior. How then should we conceive of an imaginary change in "race" which would imply

[5] Within causal approaches to algorithmic fairness, this central question is framed as the distinction between "unfair" or "fair" causal paths, or between "unresolved" or "resolved" causal influences, leading from "race" to "score" [6, 24].

[6] Assuming a specific (social, non-biological, constructivist, non-reductionist) understanding of "race", some authors claim that the counterfactual notion of some person having (merely) a different race is incomprehensible in the first place and useless for debates on discrimination [23, 26].

discrimination if accompanied by a change in "score"? Again, this can be taken as an inherently normative question, and an adequately designed answer will be largely parallel to the above one. All those changes in other variables that can be traced back to *past or present discrimination* should be in the counterfactual picture of the person belonging to a different race. For if an algorithm's predictions changed with those variables, it would *perpetuate these discriminating effects* of historical wrongs or false data, and so be in need of debiasing corrections. Consequently, counterfactual reasoning has again brought us back to exactly those issues that already appeared pivotal for discussions of statistical fairness.

In short, whether (indirect) causal paths leading from "race" to risk scores are to be regarded as instances of bias, or which (imagined) counterfactual scenarios where "race" is switched and risk scores shift too should count as indicators of discrimination largely depends on the causes that establish these relationships in the first place. And adequate countermeasures against the workings of these correlations must ultimately be based on assessments of past or present discrimination that bring them about. At the same time, one may doubt whether it is necessary, and actually possible, to conduct these assessments down to the levels of single causal paths or even individual persons, as approaches of causal reasoning and counterfactual fairness tend to suggest. General policies that need to be established in fighting algorithmic discrimination in the penal sector may well be allowed to, and may ultimately have to, restrict themselves to a critical awareness of the sociological impact of discriminatory practices on correlations between "race" and scores, without spelling out their psychological mechanisms through concrete traits in specific persons.

3. DISCURSIVE FAIRNESS

3.1 Another Problem in COMPAS

It must be emphasized that nothing in the above discussion of statistical fairness is unique to AI predictions. Human predictions, whether in the forensic sector or in other social spheres, are affected by the same problems, i.e. the basic plurality of fairness measures and the general impossibility of their simultaneous fulfilment. But COMPAS opens up another problem, which we might phrase as a problem of discursive fairness. This problem, in contrast to statistical fairness, is specific to AI predictions, or more precisely to human decisions based on AI predictions.

Let us imagine that there were no statistical fairness problems in COMPAS, i.e. no true or spurious correlations between race and recidivism, no differing error rates for blacks and whites, and perfect matches of positive predictive values. Even if this were the case, we still might question the use of COMPAS predictions in court, pointing to issues which are commonly framed as "black box problem", "lack of transparency", or "right to explanation" [8, 10, 22]. But it is possible to reconstruct these issues once more under the heading of discrimination. This will help to make more explicit what the black box problem amounts to, what kind of transparency is required, and what rights are at stake, in the given penal context.

3.2 Definition of Discrimination

To see this it will be useful to start off from the following working definition of (wrongful) discrimination, which aims to capture the essential factual and normative dimensions of the concept [1]: (Wrongful) discrimination consists in *differentiating* between persons, particularly *disadvantaging* certain persons belonging to *salient groups*, for no *relevant reason*, notably *just because* of their belonging to a *salient group*. Focusing on "salient groups" brings in a *historical dimension*. More accurately, it is the history of a given society which determines whether some social group has been exposed to widespread disadvantaging, so that the corresponding feature marks out a vulnerable subpopulation, defined e.g. by race, gender, ethnicity, or religion. Clarifying what amounts to a "relevant reason" opens up a *contextual dimension*. More precisely, the question whether something is a relevant reason or not will depend on the social system and the corresponding decision processes envisaged, e.g. university admissions, loan granting, school tests, or job hiring.

Admittedly, within the given context of crime prediction for penal decision-making in court hearings there may be different accounts of what a "relevant reason" for denying a defendant bail, probation or parole must ultimately amount to. The alternative imperatives of focusing either on the avoidance of unnecessary imprisonment or on the protection of the general public leave this issue largely open (see Section 2.2). We may have been able to make some statements concerning statistical fairness, pointing out which imbalances between groups appear to be particularly troubling, given their past or present causes. But this does not give us concrete guidance on which levels of disposition to recidivism might justify waiving punishment and which might not.

Fortunately, however, for our present purpose we need not make any definite statements on these matters. What is important for the current discussion is merely that, whatever precise standard of penal justice we may subscribe to, if we decide to deny a defendant bail, probation or parole, we need to justify this decision by providing reasons for it. This is a basic demand of discursive fairness. For if we cannot provide reasons, we will have an instance of wrongful discrimination, differentiating between persons for no relevant reason (see definition above).

Note that according to this perspective, the focus is no longer on AI predictions and their statistical qualities, which may need to be adjusted, either by a judge or by a programmer. Rather, the focus is on human decision-making based on AI predictions, and on the specific justificatory demands that social decisions concerning other persons' fates entail. In our context, it is the specific discursive setting of a penal trial that will determine what may count as a relevant reason. As stated above, we will not need to establish substantial sets of reasons that are valid in this regard, but we can restrict ourselves to narrowing down the formal types of reasons that might serve in such justifications.

3.3 The Core Question

In order to address this problem of discursive fairness in AI-based penal decision-making, we need to answer another core question: *What reasons must a judge provide for her decision when denying a defendant bail, probation or parole?* Such a decision amounts to differential treatment of the defendant, compared to other defendants who were granted these advantages, and so the judge must give relevant reasons for this differential treatment, in order not to generate a clear instance of straightforward discrimination. However, when basing her decision on COMPAS risk scores, a couple of answers that the judge might want to produce are clearly insufficient for this purpose.

First attempt: "I did not decide, but COMPAS did!" – This answer is plainly wrong: the judge signed the judgment, and this very procedure of signing the judgment is what making a judicial decision consists in. Besides, if the answer was true, it would be bad news for the judge: judges are paid to make these decisions,

and if she could demonstrate that she did not make the decision, she would have to return her salary.

Second attempt: "I did decide, and my reason was the COMPAS risk score!" – This answer is beside the point: we did not ask for the subjective reason that may have prompted the judge's decision, i.e. a psychological explanation of her action. What we require as an answer is something entirely different: we ask for the objective reason that may account for her decision's adequacy, i.e. a legal justification of her action.

Third attempt: "But COMPAS is very reliable. Accordingly, its risk score should be taken as an objective reason!" – As a matter of fact, COMPAS's reliability is not all too impressive: a positive predictive value of around 60%, implying that only 6 out of 10 of COMPAS's predictions turn out to be true, is not that good.[7] But even if it were, or if we concluded that better estimates for future human criminal behavior are not available, either because of epistemic limits to such foreknowledge or because of ontological indeterminacies in human behavior, the answer would still be misplaced: COMPAS's position in the judicial process is comparable to that of an expert witness, and so positive evidence needs to be provided for its current recommendations, beyond just pointing to its past performances and general reliability.

Fourth attempt: "I know important details of COMPAS's structure, training, working, mechanism, including its problems, such as the diverging FPRs and the mathematical incompatibility of different fairness measures. Against this background, I do have an objective reason to take its predictions into account!" – Indeed, this is not true: the basic structure and subsequent training of COMPAS is a commercial secret of Northpointe, not revealed to the judges or the public.[8] But even if these facts about COMPAS were available, comparing its position to that of an expert witness once again demonstrates why the answer is not satisfying: we do not want details about an expert witness's brain structure, school training, mental processes, or reasoning styles either, but rather, we want a reason why some person is not eligible for bail, probation or parole.

3.4 Basic Requirements of Fair Trials

The four responses sketched in the preceding section are all flawed. But they bring us closer to what would in fact be required for something to count as a relevant reason in a court setting, when basing a denial of bail, probation or parole on a prediction of future criminal behavior. Certainly, some explanation of this prediction is necessary for a justification of the decision. But how much and what kind of explanation is demanded?

This problem must be addressed with regard to the discursive standards of a fair trial. Against this background, there seem to be two questions that the judge has to answer: First, which *feature of the defendant* makes her suggest that the defendant might reoffend (and thus is not eligible for bail, probation or parole)? Second, what *psychosocial regularity or causal mechanism* is assumed in this prediction (and thus in her decision)?

These questions must be answered within the discursive setting of a criminal proceeding. This is because the defense must be able to challenge the decision, and this can happen in two essential ways: The defense may either provide evidence that the defendant *does not have* the feature in question (by calling a witness, by submitting relevant documents, etc.). Or the defense may provide evidence that the regularity or the mechanism presumed *does not hold* (by hearing an expert, by pointing to recent research, etc.).

As a consequence, the following suggests itself as a first approximation to the above problem: a relevant reason required to prevent wrongful discrimination in a court decision on bail, probation or parole must specify *(i) decisive features of the defendant* presumed to make future offences from his side sufficiently probable, and *(ii) empirical regularities or scientific mechanisms* assumed to support this predictive verdict. This is a formal requirement which must be met independently of substantial debates on the factual reliability of both pieces of evidence, or the normative impact that they should have: We may argue about whether a defendant has the feature in question, or whether it is indeed predictive of criminal behavior. We may debate whether these facts should suffice to foreclose bail, probation or parole. But this discussion can only proceed, within a fair trial, when the two pieces of information are provided as reasons for the decision.

So this is the kind of "transparency" that is required for algorithmic predictions in penal settings. Its content is specified with regard to the discursive setting of a fair trial. COMPAS does not fulfil this criterion, because of its "black box" character, and so it violates a defendant's "right to explanation", in a very clear sense. In particular, the problem is not the lacking transparency of COMPAS's basic construction, owed to its commercial background, but the lacking transparency of COMPAS's concrete predictions, due to their unspecified references.

Note that this is a qualitative difference to human experts, such as psychologist consultants. To be sure, human experts, when appearing in court, may make risk predictions based on specialist theories that are not fully intelligible to judges or defense attorneys. However, they are still able to, and they will be asked to, state clearly the features of the case they consider paramount for their assessments and the regularities of behavior they assume in their prognoses. This is not just to let them demonstrate the epistemic quality of their predictions, but rather to enable others to challenge their predictions through targeted counter-evidence.

3.5 Diverging Opinions

The above argument is contrary to *State vs. Loomis*, a famous judgment of the *Supreme Court of Wisconsin* (July 13, 2016) in which it was decided that the due process rights of (white) defendant Eric Loomis were not infringed by the use of COMPAS risk scores in the trial against him [33]. In particular, his right to be sentenced based upon accurate information, including his opportunity to assess this accuracy and challenge its validity, was

[7] Indeed, much simpler algorithms, referring to considerably fewer items (age, sex, and number of past convictions) and working in a completely transparent way (through a decision tree), seem to perform as well as COMPAS in terms of reliability [2]. This fact might provide courts with additional doubts concerning whether it is worth carrying the costs of this commercial product and accepting its inherent lack of transparency.

[8] This fact, along with the problem of mislabeling (re-)arrests as (re-)offences (see Section 2.3) and the lack of transparency in results (see Section 3.4), seems to constitute a major knock-out argument against the use of COMPAS in its present form [22]. Public institutions in general, and penal courts in particular, should not accept this policy and rather demand that private firms, if they want to sell their services to public authorities, fully disclose the structures and workings of their products.

not considered to have been compromised in his trial. In its justification, the *Supreme Court of Wisconsin* argued that COMPAS used individual information on Loomis himself (collected from his criminal file and personal interviews) and proved statistically reliable in published validation studies (notwithstanding certain limitations and race correlation issues). Both pieces of evidence could be checked by Loomis so that his rights to due process were not infringed.

However, both levels of information are far too unspecific in order to grant the defense adequate opportunity to mount a legal challenge. The Loomis side needs to know, first, *which of the 137 items* were used in his case (and to what extent), and, second, *which empirical regularity* was assumed to obtain in the prediction (and how it was supported). It is only when provided with this information that the defense can launch a targeted challenge of the impending decision. So contrary to the opinion of the *Supreme Court of Wisconsin*, the use of COMPAS risk scores constitutes a clear violation of procedural justice.

4. CONCLUSION

Both accounts, statistical fairness and discursive fairness, allow for no general answer concerning what definition of fairness we should apply, or what standards of fairness we should adhere to. Adequate answers can only be approached with close regard to the historical facts we face and the concrete systems we are talking about. In discussing *statistical fairness* we need to look into the *empirical causes* for perceived correlations between race and recidivism, in order to establish which aims are paramount in debiasing AI predictions, i.e. which fairness measures are most relevant. In discussing *discursive fairness* we need to specify the *relevant reasons* in judicial proceedings on bail, probation or parole, which are determined by the justifications that must be provided for court decisions based on AI predictions in a fair trial.

Finally, both dimensions must be brought into close contact, as it is exactly the combination of the two fairness dimensions which may help us to avoid plainly insufficient accounts. As mentioned above, focusing exclusively on statistical fairness may end up in devising algorithms that satisfy fairness standards by issuing absurd predictions (equalizing predicted rates by way of randomization, equalizing false positives rates through deliberate misclassification of harmless persons). Taking into account discursive fairness may safeguard against these obvious malpractices (requiring justification of penal decisions precludes tossing coins, or detaining innocuous people). So being forced to mark out decisive features of persons and assumed regularities in predictions in the name of discursive fairness may prevent misguided versions of statistical fairness.

The considerations in this paper have closely referred to the specifics of discrimination against black people in the US, and the basic tenets of fair trials. The concrete statements arrived at are not directly transferrable to other settings and systems, such as gender discrimination in job hiring. However, similar observations might hold for these alternative applications as well. Monitoring empirical causes for observed correlations and defining relevant reasons for justifiable decisions may prove to be of paramount importance in many fields of AI-based decision-making.

5. ACKNOWLEDGMENTS

This work was carried out as part of the project *Bias and Discrimination in Big Data and Algorithmic Processing – BIAS* (www.bias-project.org), funded by Volkswagen Foundation. Many thanks to Markus Ahlers, Philippe van Basshuysen, Uljana Feest, Mathias Frisch, Caroline Gentgen, Christian Heinze, Jan Horstmann, Tina Krügel, Wolfgang Nejdl, Eirini Ntoutsi, Bodo Rosenhahn, Arjun Roy and Jannik Zeiser for numerous discussions, and to Lucie White for careful checking of the final manuscript.

6. REFERENCES

[1] Altman, A. Discrimination. The Stanford Encyclopedia of Philosophy (Winter 2020 Edition), ed. E.N. Zalta. https://plato.stanford.edu/archives/win2020/entries/discrimination.

[2] Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., and Rudin, C. Learning Certifiably Optimal Rule Lists for Categorical Data. Journal of Machine Learning Research 18, 234 (2018), 1–78.

[3] Angwin, J., Larson, J., Mattu, S., and Kirchner, L. Machine Bias. There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica, May 23, 2016. www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[4] Balko, R. There's overwhelming evidence that the criminal justice system is racist. Here's the proof. The Washington Post, June 10, 2020, www.washingtonpost.com/graphics/2020/opinions/systemic-racism-police-evidence-criminal-justice-system/#Policing.

[5] Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. Fairness in Criminal Justice Risk Assessments: The State of the Art. University of Pennsylvania, Department of Criminology, Working Paper No. 2017-1.0, May 25, 2017. https://crim.sas.upenn.edu/sites/default/files/2017-1.0-Berk_FairnessCrimJustRisk.pdf.

[6] Chiappa, S., and Isaac, W.S. A Causal Bayesian Networks Viewpoint on Fairness. arXiv:1907.06430v1, July 15, 2019.

[7] Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. arXiv:1703.00056v1, February 28, 2017.

[8] Citron, D.K. Technological Due Process. Washington University Law Review 85, 6 (2008), 1249–1313.

[9] COMPAS Risk Assessment. www.documentcloud.org/documents/2702103-Sample-Risk-Assessment-COMPAS-CORE.html.

[10] Crawford, K., and Schultz, J. Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms. Boston College Law Review 55, 1 (2014), 93–128.

[11] Dieterich, W., Mendoza, C., and Brennan, T. COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity. Northpointe, July 8, 2016. www.documentcloud.org/documents/2998391-ProPublica-Commentary-Final-070616.html.

[12] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness Through Awareness. arXiv:1104.3913v2, November 29, 2011.

[13] Equivant. Practitioner's Guide to COMPAS Core. April 4, 2019. www.equivant.com/practitioners-guide-to-compas-core.

[14] Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. Certifying and removing disparate impact. arXiv:1412.3756v3, July 16, 2015.

[15] Fish, B., Kun, J., and Lelkes, Á.D. A Confidence-Based Approach for Balancing Fairness and Accuracy. arXiv:1601.05764v1, January 21, 2016.

[16] Friedler, S.A., Scheidegger, C., and Venkatasubramanian, S. On the (im)possibility of fairness. arXiv:1609.07236v1, September 23, 2016.

[17] Fullinwider, R. Affirmative Action. The Stanford Encyclopedia of Philosophy (Summer 2018 Edition), ed. E.N. Zalta. https://plato.stanford.edu/archives/sum2018/entries/affirmative-action.

[18] Galhotra, S., Brun, Y., and Meliou, A. Fairness Testing: Testing Software for Discrimination. Proceedings of 11th Joint Meeting of the European Software Engineering Conference and the ACM SIGSOFT Symposium on the Foundations of Software Engineering, Paderborn, Germany, September 4–8, 2017 (ESEC/FSE '17), 498−510. https://doi.org/10.1145/3106237.3106277.

[19] Glymour, B., and Herington, J. Measuring the Biases that Matter. The Ethical and Casual Foundations for Measures of Fairness in Algorithms. Proceedings of the Conference on Fairness, Accountability, and Transparency, Atlanta (GA), USA, January 29–31, 2019 (FAT* '19), 269–278. https://doi.org/10.1145/3287560.3287573.

[20] Hardt, M., Price, E., and Srebro, N. Equality of Opportunity in Supervised Learning. arXiv:1610.02413v1, October 7, 2016.

[21] Hutchinson, B., and Mitchell, M. 50 Years of Test (Un)fairness: Lessons for Machine Learning. Proceedings of the Conference on Fairness, Accountability, and Transparency, Atlanta (GA), USA, January 29–31, 2019 (FAT* '19), 49–58. https://doi.org/10.1145/3287560.3287600.

[22] Joh, E.E. Feeding the Machine: Policing, Crime Data, & Algorithms. William & Mary Bill of Rights Journal 26, 2 (2017), 287–302.

[23] Kasirzadeh, A., and Smart, A. 2021. The Use and Misuse of Counterfactuals in Ethical Machine Learning. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT'21). ACM, New York, NY, USA, 228–236. DOI: https://doi.org/10.1145/3442188.3445886.

[24] Kilbertus, N., Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B. Avoiding Discrimination through Causal Reasoning. arXiv:1706.02744v2, January 21, 2018.

[25] Kleinberg, J., Mullainathan, S., and Raghavan, M. Inherent Trade-Offs in the Fair Determination of Risk Scores. arXiv:1609.05807v2, November 17, 2016.

[26] Kohler-Hausmann, I. Eddie Murphy and the Dangers of Counterfactual Causal Thinking About Detecting Racial Discrimination. Northwestern University Law Review 113, 5 (2019), 1163–1227.

[27] Kusner, M., Loftus, J., Russell, C., and Silva, R.: Counterfactual Fairness. arXiv:1703.06856v3, March 8, 2018.

[28] Larson, J., Mattu, S., Kirchner, L., and Angwin, J. How We Analyzed the COMPAS Recidivism Algorithm. ProPublica, May 23, 2016. www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

[29] Lewis, D. Causation. The Journal of Philosophy 70, 17 (1973), 556−567.

[30] Makhlouf, K., Zhioua, S., and Palamidessi, C. On the Applicability of Machine Learning Fairness Notions. SIGKDD Explorations 23(1), ACM, 2021.

[31] Northpointe. Practitioners Guide to COMPAS. August 17, 2012. www.northpointeinc.com/files/technical_documents/FieldGuide2_081412.pdf.

[32] Pearl, J. Causality. Models, Reasoning, and Inference. 2nd ed., New York: Cambridge University Press 2009.

[33] State v. Loomis. 881 N.W.2d 749 (2016) 749 (Wis. 2016).

[34] Verma, S., and Rubin, J. Fairness Definitions Explained. Proceedings of the International Workshop on Software Fairness, Gothenburg, Sweden, May 29, 2018 (FairWare '18), 1–7. https://doi.org/10.1145/3194770.3194776.

[35] Zafar, M.B., Valera, I., Gomez Rodriguez, M., and Gummadi, K.P. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. arXiv:1610.08452v2, May 8, 2017.

About the author:

Dietmar Hübner is Professor of Practical Philosophy, particularly Ethics of Science at Leibniz University Hannover. His main research is in general ethics, applied ethics, and political philosophy. He holds a diploma (University of Bonn) and an M.Phil. (University of Cambridge) in physics. He earned his Ph.D. in philosophy with a dissertation on decision theory and philosophy of history, and completed his habilitation in philosophy with a book on metaphorical accounts in distributive justice (both University of Bonn). Dietmar Hübner is principal investigator in the interdisciplinary research project *Bias and Discrimination in Big Data and Algorithmic Processing – BIAS* (www.bias-project.org), funded by Volkswagen Foundation.