AI & SOCIETY
https://doi.org/10.1007/s00146-022-01441-y

ORIGINAL ARTICLE

Algorithmic fairness through group parities? The case of COMPAS-SAPMOC

Francesca Lagioia[1,2] · **Riccardo Rovatti**[3] · **Giovanni Sartor**[1,2]

Received: 30 July 2021 / Accepted: 24 March 2022

Abstract
Machine learning classifiers are increasingly used to inform, or even make, decisions significantly affecting human lives. Fairness concerns have spawned a number of contributions aimed at both identifying and addressing unfairness in algorithmic decision-making. This paper critically discusses the adoption of group-parity criteria (e.g., demographic parity, equality of opportunity, treatment equality) as fairness standards. To this end, we evaluate the use of machine learning methods relative to different steps of the decision-making process: assigning a predictive score, linking a classification to the score, and adopting decisions based on the classification. Throughout our inquiry we use the COMPAS system, complemented by a radical simplification of it (our SAPMOC I and SAPMOC II models), as our running examples. Through these examples, we show how a system that is equally accurate for different groups may fail to comply with group-parity standards, owing to different base rates in the population. We discuss the general properties of the statistics determining the satisfaction of group-parity criteria and levels of accuracy. Using the distinction between scoring, classifying, and deciding, we argue that equalisation of classifications/decisions between groups can be achieved thorough group-dependent thresholding. We discuss contexts in which this approach may be meaningful and useful in pursuing policy objectives. We claim that the implementation of group-parity standards should be left to competent human decision-makers, under appropriate scrutiny, since it involves discretionary value-based political choices. Accordingly, predictive systems should be designed in such a way that relevant policy goals can be transparently implemented. Our paper presents three main contributions: (1) it addresses a complex predictive system through the lens of simplified toy models; (2) it argues for selective policy interventions on the different steps of automated decision-making; (3) it points to the limited significance of statistical notions of fairness to achieve social goals.

Keywords Fairness · Group-parity · Classifiers · COMPAS · Automated decision · Affirmative action

1 Introduction: predictions and decisions in machine learning

As the use of machine learning (ML) methods in decision-making processes has become pervasive, having the potential to significantly affect human lives, fairness concerns have grown (Barocas et al. 2017; Mayer-Schönberger and Ramge 2018; Hildebrandt 2020; Vinuesa et al. 2020). These concerns have spawned many contributions aimed at identifying and measuring unfairness in decision-making and at proposing remedies (Žliobaitė 2017; Zafar et al. 2017; Joseph et al. 2016; Hajian and Domingo-Ferrer 2012; Hellman 2020; O'Neil 2016; Kusner et al. 2017).

Most contributions have focused on the *outcomes* of machine learning systems that disparately affect sensitive groups (e.g., groups identified by race or gender).

✉ Francesca Lagioia
francesca.lagioia@eui.eu

Riccardo Rovatti
riccardo.rovatti@unibo.it

Giovanni Sartor
giovanni.sartor@unibo.it

[1] Cirsfid-Alma AI, Law Department, University of Bologna, Bologna, Italy

[2] Law Department, European University Institute, Florence, Italy

[3] Department of Electrical, Electronic, and Information Engineering "Guglielmo Marconi", University of Bologna, Bologna, Italy

Published online: 28 April 2022

On one hand, such outcomes, or rather the decisions based on them, have been evaluated by applying categories of anti-discrimination laws, such as the distinction between disparate treatment (also called direct discrimination) and disparate impact (also known as indirect discrimination). In the first case, the detrimental outcome is based on prohibited features. In the second case, such an outcome is based on apparently neutral features, criteria, and practices (Barocas and Selbst 2016; De Vos 2020) the consideration of which disproportionately affects a protected group, without an acceptable rationale (see Friedman and Nissenbaum 1996).

On the other hand, some abstract criteria and metrics have been developed to determine when the outcomes of machine learning systems affect individuals and groups differently (Angwin et al. 2016; Dieterich et al. 2016; Hardt et al. 2016; Chouldechova 2017; Kleinberg et al. 2016; Berk et al. 2018). According to these criteria, a decision process is called "fair," under a particular criterion, if its outcomes, relative to the groups being considered, equally satisfy certain statistical properties. For instance, a classifier is deemed fair under the statistical parity criterion if it provides an equal proportion of positive and negative predictions across all groups. These fairness notions depart from the concepts of fairness so far used in social and philosophical disciplines (see Rawls 2001; Rescher 2002). They point to group differences, which may have different grounds, depending on the predictive system's biased functioning or on differences in the underlying populations.

The fairness analyses of decisions based on machine learning usually do not distinguish the different steps involved in a decision-making process. Consequently, they are unable to identify the ways in which decision-making processes can be improved through specific interventions. To fill this gap, we distinguish the following steps: assigning a predictive score, linking a classification to that score, and taking a decision (i.e., selecting actions) based on that classification.

The first step fundamentally consists in an epistemic determination: it provides a factual assessment, usually expressed through a numeric score, that approximates the likelihood that things are in a certain way, or that they will evolve in a certain direction. On the contrary, both classification and decision-making involve practical judgements based on the epistemic determination provided by the score: they are geared toward achieving the goals (economic, ethical, political, etc.) driving the decision-making process, including a fair treatment of individuals. By separating out epistemic and practical determinations, it may be possible to achieve a less controversial and more focused assessment of automated decisions. In some cases, we may agree that the system's predictions are epistemically faulty; in other cases, while agreeing on their epistemic correctness, we may disagree on the appropriateness of the following classifications and decisions, if we differ about the political-ethical values at stake or about their relative weight.

In the following, we first distinguish the three stages mentioned above, i.e., (1) the computation of a probabilistic score, (2) the classification based on the score, and (3) the decision based on the classification. Then we focus on the use of predictive technologies to evaluate the risk of recidivism. We briefly introduce the COMPAS system, the Loomis case in which it was challenged, and the ensuing debate on COMPAS's fairness.

To provide clearer insights and to support appropriate generalizations, we provide a simplified version of COMPAS, which we call SAPMOC I, and assess its outcomes using some standards for group parity. We show how a system equally accurate for two groups may fail to comply with some parity standards owing to different base rates in the population.

We discuss the general properties of the statistics determining the satisfaction of parity criteria and levels of accuracy. By expanding our toy example into SAPMOC II, we show that equalization of classifications/decisions between groups can be achieved by way of group-dependent thresholding. We discuss contexts in which this approach may be meaningful, and useful in pursuing policy objectives in light of socio-political preferences. We argue that predictive systems should be built in such a way that relevant preferences can be transparently introduced by human decision-makers.

Our paper presents three main contributions: (1) it addresses a complex predictive system (COMPAS) through the lens of simplified toy models (SAPMOC I and SAPMOC II); (2) it argues for selective policy interventions on the different steps of automated decision-making (i.e., scoring, classifying, deciding); (3) it points to the limited significance of statistical notions of fairness to achieve social goals in different contexts.

2 The anatomy of decisions

As noted by Agrawal et al. (2018), machine learning systems can be viewed as "prediction machines-" In comparison to human decision-making, they provide in many domains for more precise and cheaper predictions, and consequently lead to a much greater number of predictions being made. Predictions may concern both the existence of factual preconditions for engaging in certain actions, as well as the expected outcomes of such actions. As examples in the medical domain, consider the prediction that a patient has a certain pathology as opposed to the prediction that a certain therapy will be effective. As examples in the justice domain, compare the prediction that a certain individual will recidivate (reoffend), and the prediction that a correctional measure will be effective for his or her social reintegration.

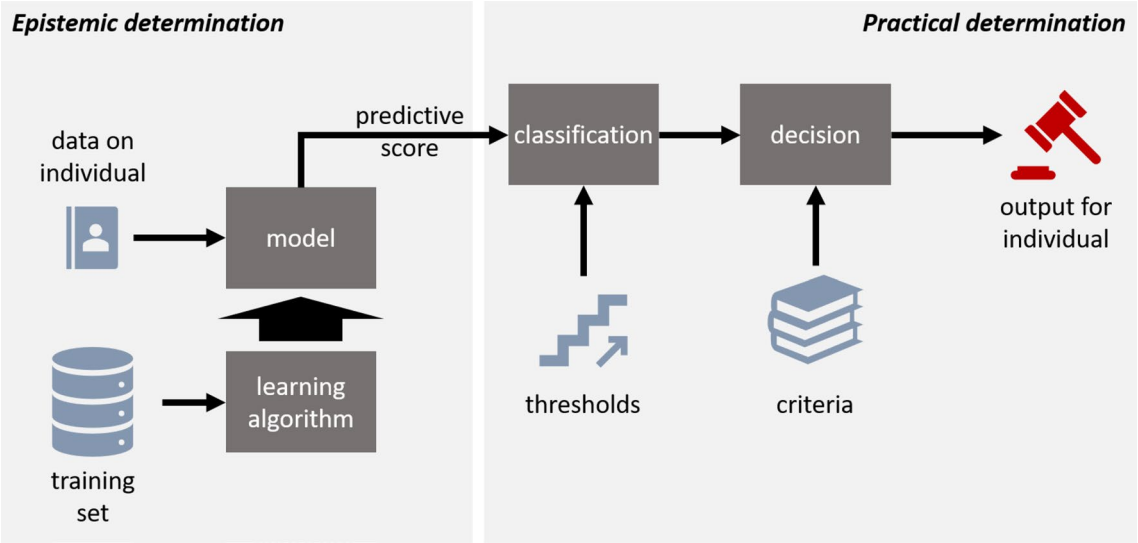

Fig. 1 The anatomy of decisions

It is important to remark that predictions only are one component in a larger process. Decision-making is not limited to predicting but also requires specifying the goals to be pursued, identifying the applicable ethical or legal constraints, evaluating the predicted consequences of alternative courses of action, and selecting the action best suited to goals and constraints.

As shown in Fig. 1, we distinguish three main steps in a decision-making process supported by predictive algorithms: scoring, classifying and deciding.

2.1 Predictive scoring

The first step is predictive scoring, which consists in assigning a score to an entity. The score expresses the likelihood that the entity has the predicted property (see Citron and Pasquale 2014). Depending on the domain, different target properties can be predicted. For instance, when the task is to determine whether an industrial product may be defective, the score expresses the likelihood that the product is indeed faulty. Where the risk of fraud is being predicted, the score indicates the likelihood that a transaction will indeed be fraudulent. This first step, as we shall argue, should be fundamentally based on epistemic considerations, namely, on getting scores that most accurately reflect the likelihood of the target properties being present.

As shown in Fig. 1, where machine learning methods are adopted, the score is the outcome delivered by a model (a learned algorithm) that is constructed by another algorithm (the learning algorithm). Supervised learning is based on a training set, i.e., a set of examples, each linking the values of certain features (the predictors) in a particular case to the value of the feature being predicted (the label) in the same case. For instance, in medical diagnosis, each example may link the features of particular patient (e.g., medical history and scans) to the pathologies by which the patient is affected. In the case of recidivism, each example may link the features of a past offender (e.g., criminal record and psychological traits) to the offender's behavior after release.

2.2 Classification

The second step in the pipeline—classification—is constitutive rather than descriptive. Classification is not meant to identify "objective" features of the entities to which it is applied, but rather to provide triggers for action. It must be considered in connection with the decisions that may be taken or considered depending on the labels ascribed through classification.

More to the point, classification mediates threshold scores and decisions, as shown in Fig. 2. For instance, by classifying entities (within a certain score interval) as having a high, medium, or low likelihood of possessing the target feature, we anticipate the way in which such entities are going to be treated. For instance, assume that a hospital has a policy under which high-risk patients are to undergo certain medical tests. By choosing to classify as high-risk all patients whose score indicates a greater than 20% chance of having a certain pathology, we determine what patients will be subject to these tests. Similarly, assume that a policy exists under which transactions with a high risk of fraud will be addressed by blocking the credit cards of the parties involved. By choosing to classify as high-risk only those transactions that have a greater than 85% chance of being fraudulent, we determine what transactions will trigger this measure.

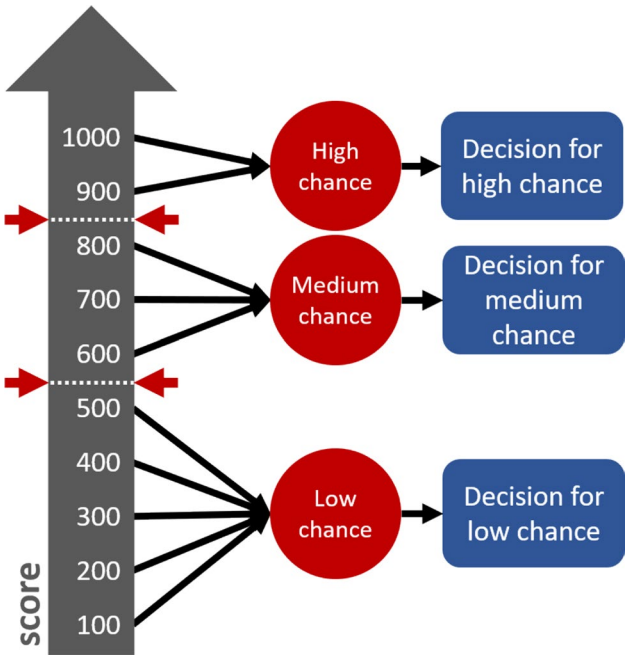

Fig. 2 General classification

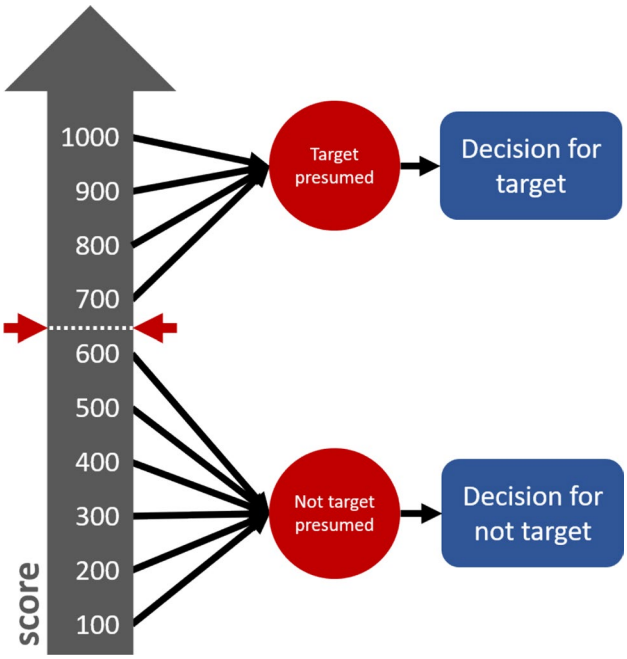

Fig. 3 Binary classification

As these examples show, threshold-setting and the consequent classifications are non-neutral: they are rather goal- and cost-driven. In the medical-testing example, the threshold was set low, given the high cost of failing to detect an instance of disease. In the fraud-detection example, a high threshold was set in view of the cost, in terms of reduced customer satisfaction, of blocking the credit card of an innocent user.

Figure 3 shows the function of score thresholds: they separate those cases that are to be treated as having the target feature from those cases that are to be treated as missing such feature. Thus, thresholds can be said to introduce a kind of presumption relative to the possession of the target feature, based on cost-effectiveness analyses combining the expected benefits and the costs of interventions. All entities above the threshold are considered to be positive, even if only some of them, i.e., those having the target feature, will be true positives, the others being false positives. Similarly, all entities below the threshold are considered to be negative, even if only some of them, i.e., those missing the target feature, will be true negatives, the others being false negatives. For instance, in tools assessing the risk of recidivism, setting a certain threshold for high risk will entail that all those above the threshold will be considered future recidivists (including false positives) and all those below it will be considered as future non-recidivists (including false negatives).

2.3 Decision

The final step is decision. Decision may be entrusted to a human decision-maker, who may take other situational aspects into account, or it may be automatically linked to classification by means of a computable rule. As an example of a human decision based on an automated classification/score, consider a physician who evaluates a diagnosis suggested by a predictive system and decides the appropriate treatment accordingly. Similarly, a judge may consider the recidivism prediction and risk-classification of defendants in deciding on a correctional or treatment program (as in the COMPAS case, which will be extensively considered in what follows). As an example of an automated decision, consider a loan application which is automatically rejected by a computer system, since the applicant is classified as high risk. Similarly, a product classified as defective can be automatically discarded.

3 The COMPAS system and the Loomis case

In this section, we examine COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), an actuarial risk and need-assessment instrument widely used in the United States. COMPAS is deployed in the criminal justice system for evaluating defendants' risk profiles: risk of recidivism, risk of violence, and risk of failure to appear in court.

The assessments made by COMPAS are taken into account by judges in deciding whether to grant the benefit of parole/probation. Risk is assessed through statistical algorithms and quantified into risk scores. Such scores are computed on the basis of multiple data points, which include static-historical factors (such as criminal history, age of first arrest, criminal associates) and dynamic-criminogenic factors (such as residential stability, employment status, community ties, substance abuse, social inclusion and relationships, and family status), as well as answers to 137 multiple-choice questions. COMPAS has two primary risk models: General Recidivism and Violent Recidivism. The General Recidivism Risk Scale is used to predict new offences. The Violent Recidivism Risk Scale focuses on the probability of violent crimes, i.e., murder, manslaughter, rape, robbery, and aggravated assault. COMPAS scale scores are transformed into decile scores by dividing these scores into ten equally sized groups. In particular, scores in deciles from 1 to 4 are labelled "Low" risk; from 5 to 7 "Medium"; and from 8 to 10 "High."

3.1 A legal challenge: the Loomis case

The use of COMPAS has been widely debated in the wake of the case of *Loomis v. Wisconsin*.[1] Eric Loomis was charged with driving a stolen vehicle he used in a shooting and fleeing from police. Before deciding the case, the Circuit Court of Wisconsin ordered a presentencing investigation in part based on the COMPAS assessment. As a result, Loomis was classified as being at high risk of reoffending (Brennan et al. 2009) and was sentenced to 6 years of imprisonment and 5 years of extended supervision.

Loomis appealed the Court's decision, arguing that the Court's reliance on COMPAS violated his due process rights, on the following grounds.[2] First, COMPAS does not disclose how risk scores are computed. This lack of transparency prevents defendants from challenging the scientific validity and accuracy of such scores. Second, COMPAS reflects race and gender biases. In particular, black men have a higher likelihood of being mistakenly predicted to reoffend, and females are assigned a lower risk score, all the rest being equal. Third, the system's predictions are based on statistical correlations. Thus, the court's use of COMPAS infringes both the right to an individualized sentence and the right to be sentenced on accurate information.[3] The Supreme Court of Wisconsin rejected the defendant's arguments. Regarding the COMPAS system's opacity and accuracy, the court held that even though Loomis could not review and challenge COMPAS computations, he could still review and challenge the resulting risk scores and the factors on which they were based, some being publicly available and other being provided by the defendant.

The Supreme Court of Wisconsin denied that COMPAS discriminates against men, stating that the use of gender as a factor in risk assessment serves the non-discriminatory purpose of promoting accuracy. Regarding the race discrimination issue—i.e., the allegation that COMPAS systematically attributes higher risk scores to black offenders than to white ones—the Wisconsin Supreme Court merely highlighted the importance of adequately informing COMPAS users about the related debate.

Finally, the Court admitted that COMPAS statistical algorithms are based on generalizations, since the likelihood that a person reoffends is computed on the basis of the past behavior of similar individuals. However, the court explained that COMPAS is merely meant to enhance the evaluation of judges, who should weigh *all* the available evidence in determining an individualized program appropriate to the defendant.

3.2 A statistical challenge: the ProPublica study

The *Loomis* case has been widely reported and debated in the scholarly literature (Angwin et al. 2016; Flores et al. 2016) and beyond (Angwin et al. 2016; Yong 2018; Liptak 2017; Tashea 2017), which has challenged the accuracy and fairness of COMPAS. In 2016 ProPublica, a nonprofit organization specialized in investigative journalism, published an extensive study (Angwin et al. 2016) based on 11,757 defendants in Broward County, Florida, assessed by COMPAS in 2013 and 2014. The study compared the recidivism risk rates predicted by COMPAS with the actual recidivism rates of defendants within 2-year span so as to determine the extent to which COMPAS predictions come true for different race-based groups (black, white, Hispanic, Asian, and Native American).

On this basis, ProPublica raised several criticisms, among which the following. Firstly, it argued that COMPAS is inaccurate: in many cases individuals classified as high risk did not reoffend, while those flagged as medium or low risk committed new crimes. COMPAS correctly predicted recidivism 61% of the time, and it only correctly predicted violent recidivism in 20% of cases (Larson et al. 2018).

Secondly, and most importantly, ProPublica claimed that COMPAS was racially unfair. The average probability to be predicted at a high risk of recidivism was much higher for blacks than for whites . Furthermore, the proportion of black defendants misclassified as high risk (relative to the total number of blacks who did not reoffend) was much higher than the corresponding percentage of whites (45% as opposed to 23%). Conversely, white defendants were more

[1] *State v. Loomis*, 881 N.W.2d 759 (Wis. 2016).

[2] *Loomis*, 881 N.W.2d at 756.

[3] *Loomis*, 881 N.W.2d 759 (Wis. 2016).

often predicted to be less risky than they were. White reoffenders were mistakenly labeled as low risk almost twice as much as black ones (48% as against 28%) (Larson et al. 2018).

Consequently, it appeared the COMPAS's assessment were affected by racial bias. On one hand, black non-recidivists were more likely than white non-recidivists to be erroneously subjected to the detrimental consequences linked to a recidivism prediction. On the other hand, white recidivists were more likely than black recidivists to erroneously obtain the more favorable treatment for expected non-recidivists.

In 2016, scientists from Northpointe, Inc. (Dieterich et al. 2016), challenged ProPublica's report (Angwin et al. 2016), claiming that it was based on several statistical and technical errors. Especially, the report did not take into account the different base rates of recidivism for blacks and whites. The Northpointe scientists argued that COMPAS was not racially biased, since the prediction that an individual would or would not reoffend was equally correlated, for both blacks and whites, with the likelihood that the individual would actually reoffend.

Black defendants who were predicted to reoffend actually did recidivate at a slightly higher rate than their white counterparts (63% as against 59%). Similarly, white defendants who were predicted not to recidivate did not reoffend at a slightly higher rate than black defendants (71% as against 65%). These findings provided evidence of predictive parity[4] for blacks and whites in the target population.

The authors also demonstrated that both the Recidivism Risk Scale and Violent Recidivism Risk Scale were equally accurate for blacks and whites. Finally, they pointed out that COMPAS accuracy should be evaluated with respect to the accuracy of human judgments, which on average is lower than that of the system (Dieterich et al. 2016).

3.3 From COMPAS to SAPMOC: A mock predictive system

Technical contributions addressing the COMPAS system have shown that different fairness evaluations can be made by applying different group parity standards. Understanding the ethical and legal significance of these outcomes in the COMPAS case is difficult, given the high complexity of such a system. Lawyers and other non-experts in statistical analysis/machine learning are consequently puzzled and unable to take a reasoned position in the COMPAS debate, as in other issues pertaining to algorithmic fairness.

To illustrate and clarify such issues and make them accessible to a nontechnical audience, we have adopted the following methodological approach. We have defined a toy example, which we call SAPMOC (by inverting the "COMPAS" name), which exemplifies the main source of the COMPAS controversies: the application of statistical predictions to populations characterized by different base rates relative to both the predictors and the target property. This helps us to address some key points, without getting bogged down in details and complexities. We will introduce two versions of SAPMOC, a simpler version, SAPMOC I, which only uses a single binary input feature, and a more complex version, SAPMOC II, using multiple input features.

3.4 Meet SAPMOC I

Like COMPAS, SAPMOC I assesses the risk of recidivism. However, rather than 130 input features, it only uses a single binary feature, i.e., whether the defendant committed previous offences. A realistic system for predicting recidivism should consider multiple features—education, family situation, job, income level, character, etc.—though it has been argued that the functioning of COMPAS can be reproduced using only a few features (Rudin 2019). For our explanatory purposes, however, a single feature will do.

We consider a population (P) of $N^{\mathrm{P}} = 3000$ defendants, divided into two groups, the Blue (B) or the Green (G) group, such that $P = \mathrm{B} \cup \mathrm{G}$ and $\mathrm{B} \cap \mathrm{G} = \varnothing$. We denote the number of individuals in each group by $N^{\mathrm{B}} = 1500$ and $N^{\mathrm{G}} = 1500$, so that $N^{\mathrm{P}} = N^{\mathrm{B}} + N^{\mathrm{G}}$.

Only for the sake of our example, and with no link to any real context, we assume that our input feature has a high predictive capacity: defendants having a criminal record reoffend in 80% of cases, while defendants with no criminal record reoffend in 20% of cases. We also presume that 1000 individuals in the Blue group have a criminal record, while only 500 individuals in the Green group do.

Applying the general decision-making framework sketched above—where we distinguish between assigning a score, making a classification depending on whether the score is above or below a certain threshold, and deciding accordingly—we get the simplified account in Fig. 4. Note that, given the binary nature of our predictions, just two possible scores can be assigned, e.g., 800 for having a criminal record and 200 for not having it. We assume that all those with a score of 800 are classified as high risk, i.e., as (likely) reoffenders, and all those with a score of 200 are classified as (likely) non-reoffenders. Consequently, the first are denied probation, while the second are accorded it. Note, however, that the assignment of scores is here redundant, since the functioning of SAPMOC I could be adequately captured by directly linking the presence of the input feature to the corresponding classification.

[4] A classifier exhibits "Predictive parity" if it delivers similar predictive values for two different groups (e.g., blacks and whites), such as the probability of reoffending, given a similar score for such groups.

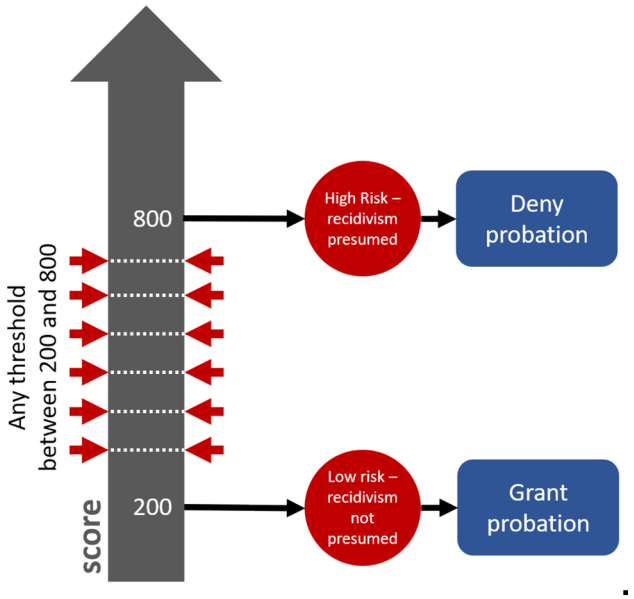

Fig. 4 Binary classification by SAPMOC I

Table 1 Real outcomes

	Record	No record
Recidivism	1200	300
No recidivism	300	1200
Total	1500	1500

Table 2 SAPMOC predictions

	Record	No record
Recidivism	1500	0
No recidivism	0	1500
Total	1500	1500

3.5 SAPMOC I's overall performance

Tables 1 and 2 reflect the correlation between criminal record (previous convictions) and recidivism across the whole population (without distinguishing Blue and Green individuals) in SAPMOC I.

SAPMOC I has classified as recidivists all individuals having a criminal record (1500). In so doing, it has erroneously classified 300 individuals (all those having a criminal record who did not recidivate). Similarly, it has classified as non-reoffenders all individuals without a criminal record (1500), thus erroneously classifying 300 individuals (all those not having a criminal record who did reoffend). As a result, SAPMOC I has incurred in 600 errors. In assessing the functioning of SAPMOC I, we must consider that errors

Table 3 SAPMOC's overall performance

	Predicted recidivist	Predicted non-recidivist
Recidivist	TP = 1200	FN = 300
Non-recidivist	FP = 300	TN = 1200

are committed by any predictor operating under real-world circumstances. The issue is whether the error rate is acceptable in the context of the application domain, in comparison with the available alternatives.

Independently of how binary predictive systems work internally, their performance can be characterized by their *confusion matrix*, which relative to SAPMOC I takes the form in Table 3.

Each row represents the instances in actual classes, while each column represents the instances in a predicted class:

1. TP is the number of true positives, i.e., those individuals for whom the prediction is positive (recidivism is predicted) and that prediction is true (the individual will reoffend),
2. FN is the number of false negatives, i.e., the individuals for which the prediction is negative (no reoffence is predicted) but that prediction is false (the individual will reoffend),
3. TN is the number of true negatives, i.e., the individuals for whom the prediction is negative (no reoffence is predicted) and that prediction is true (the individual will not reoffend), and
4. FP is the number of false positives, i.e., the individuals for whom the prediction is positive (reoffence is predicted) but that prediction is false (the individual will not reoffend).

When needed, quantities related to entire populations are indicated with a superscript ^{P}, while those referring to only one of the two sub-populations (the Blues or the Greens) are indicated with the superscripts ^{B} or ^{G}. Hence, for example, TP^{P} is the number of individuals in the global population that are predicted to be recidivist and will indeed commit further offences, while TP^{B} is the corresponding number of Blue individuals, etc.

3.6 Distribution of recidivism in different sub-populations

Consider now the Blue and the Green population separately. As above, we assume that our input feature (having criminal record) is not equally distributed across the two groups: 1000 Blue defendants have a criminal record, while only

Table 4 The base rate in the Blue and Green groups

	Blue (%)	Green (%)
Recidivists $\frac{TP+FN}{TP+TN+FN+FP}$	60	40
Non-recidivists $\frac{TN+FP}{TP+TN+FN+FP}$	40	60

Table 5 Confusion matrices in SAPMOC

Blue		Green	
TP = 800	FN = 100	TP = 400	FN = 200
FP = 200	TN = 400	FP = 100	TN = 800

Table 6 Statistical parity in SAPMOC

	Blue (%)	Green (%)
Positive class $\frac{TP+FP}{TP+TN+FN+FP}$	67	33
Negative class $\frac{TN+FN}{TP+TN+FN+FP}$	33	67

Table 7 Conditional procedure accuracy equality in SAPMOC

	Blue (%)	Green (%)
Positive class $\frac{TP}{TP+FN}$	89	67
Negative class $\frac{TN}{TN+FP}$	67	89

500 Green ones do. Within each group the criminal record is equally correlated with recidivism.

Table 4 reports the base rate of our population, i.e., the proportion of individuals who have or have not reoffended, relative to the total amount of individuals within each group. For instance, since the Blue group includes 1,000 defendants with a criminal record and 500 defendants without such a record, and the former will reoffend in 80% of cases, while the latter only in 20%, it follows that $1000 \times 80\% = 800$ individuals with a criminal record and $500 \times 20\% = 100$ without such a record will reoffend, i.e., a total of 900 Blues. Given that the Blue group includes 1,500 individuals, 900 reoffenders (i.e., positive individuals) represent 60% of Blues. As shown in Table 4, similar considerations apply to the other classes (i.e., Blue negatives, Green positives, and Green negatives). Base rates of positives and negatives is different for Blues and Greens due to the different number of previous offenders in each group.

As in the ProPublica analysis, we assume to know both the SAPMOC I' predictions and the actual outcomes. This information is shown in Table 5.

As noted above, SAPMOC I predicted that all previous offenders, e.g., all Blues with a criminal record (1000), would reoffend ($TP^B + FP^B$). Since previous offenders reoffend in 80% of cases, with regard to the Blue group predictions are correct for 800 defendants (TP^B) and incorrect for 200 (FP^B) defendants. Similarly, SAPMOC predicted that all those with no criminal records, e.g., 500 Blues, would not recidivate. Since those without a criminal history do not reoffend in 80% of cases, SAPMOC's predictions are correct in 400 cases (TN^B) and incorrect in 100 cases (FN^B). Similar considerations apply to the Green group.

3.7 Evaluating SAPMOC I's predictions under fairness criteria

In the following we examine SAPMOC I's predictions under some criteria used in recent literature on fairness in machine learning (Berk et al. 2018; Kleinberg et al. 2016). These criteria are the practical and most frequently used instances of formal nondiscrimination rules taxonomized, for example, in Barocas et al. (2021).

3.7.1 Statistical/demographic parity

Statistical parity requires the proportion of positive (recidivism) and negative (no recidivism) predictions to be equal in each group. Moving from proportions to probabilities, the probability of a positive or negative classification should be equal for all individuals in the two groups. Statistical parity is not satisfied in our model, since 67% of Blues are classified as positives, while only 33% of Greens get a recidivism prediction, as shown in Table 6 (percentages rounded to the integer).

Such a divergence depends on the different base rate in the groups. To comply with statistical parity, we should consider as non-recidivists some Blues with a criminal record or, alternatively, as recidivists some Greens without such a record. However, in both cases, SAPMOC accuracy would decrease. This would also introduce a disparate treatment, which is not justified by the individuals' features: we would, for example, consider some Blues with a criminal record as non-recidivists, even though they would have been classified as recidivists had they been Green.

3.7.2 Conditional procedure accuracy equality/equality of opportunity

According to conditional procedure accuracy equality (Berk et al. 2018), the members in each group who exhibit the same behavior should be treated equally in equal proportion. The values reported in Table 7 show that this criterion remains unsatisfied. Blue recidivists are more likely than Green ones to be correctly classified as positives, and less likely to be misclassified as negatives. The opposite is true for negative predictions, where Green non-recidivists are more likely to be correctly classified as negatives than Blues, and less likely to be wrongly classified as positives.

Table 8 False positive/false negative rates in SAPMOC

	Blue (%)	Green (%)
Positive class $\frac{FP}{FP+TN}$	33	11
Negative class $\frac{FN}{FN+TP}$	11	33

The violation of this criterion negatively affects the Blues, since they are subject to a higher number of errors leading to a detrimental treatment (i.e., mistaken recidivism ascriptions). Equality of opportunity is violated, since the proportion of Greens correctly classified as non-recidivist (TN^G), relative to all Greens who did not recidivate ($TN^G + FP^G$) is higher than the corresponding proportion for the Blues. Here, too, such differences are due to the different base rate within the groups, i.e., to the higher number of individuals with a criminal history in the Blue group. Since all individuals having a criminal record are classified as positives, and all individuals without one are classified as negatives (resulting in a 20% chance of prediction error), the related errors add up in the group that includes the higher number of previous offenders.

A similar criterion is the *false positive rate*, which reflects the frequency with which the classifier makes a mistake. In our case, that is the proportion between false positives (FP) and all the individuals who did not recidivate (FP + TN). Thus, it indicates the proportion of non-recidivists erroneously classified as recidivists. Such a ratio is higher in the Blue group, as shown in Table 8. Conversely, the false positive rate is the proportion between false positives (FN) and all the individuals who did recidivate (FN + TP). In the Green group, the ratio between false negatives (FN^G) and all the individuals who actually recidivated is higher than in the Blue group.

3.7.3 Calibration/conditional use accuracy equality

According to calibration (Berk et al. 2018; Kleinberg et al. 2016), the proportion of correct predictions should be equal for each class within each group. Thus, the ratio between true positives (TP) and the total positive predictions (TP + FP) should be the same for both groups. Similarly, the ratio between true negatives (TN) and total negative predictions (TN + FN) should be equal in the two groups. In our example, the calibration criterion is satisfied, as the predictor is uniformly correlated with the outcome, and so the proportion of correct predictions is equal to 80% in each group and class, as reported in Table 9.

3.7.4 False rate/conditional use error

The false rate criterion can be considered as the other side of calibration. It measures the proportion of erroneous predictions for each class relative to the total predictions (Binns 2020; Chouldechova 2017; Barocas and Selbst 2016). To satisfy this criterion, it is necessary that the ratio between false positives (FP) and the total amount of positive predictions (TP + FP) be equal in the two groups. The same applies to the negative class, where the ratio between false negatives (FN) and the total amount of negative predictions (TN + FN) should be equal within both groups. Table 10 shows that the criterion is satisfied in our example.

Table 9 Calibration in SAPMOC

	Blue (%)	Green (%)
Positive class $\frac{TP}{TP+FP}$	80	80
Negative class $\frac{TN}{TN+FN}$	80	80

Table 10 False rate in SAPMOC

	Blue (%)	Green (%)
Positive class $\frac{FP}{TP+FP}$	20	20
Negative class $\frac{FN}{TN+FN}$	20	20

Table 11 Treatment equality in SAPMOC

	Blue (%)	Green (%)
Positive class $\frac{FP}{FN}$	200	50
Negative class $\frac{FN}{FP}$	50	200

3.7.5 Treatment equality

According to treatment equality, the ratio between errors in positive and negative predictions should be equal across all groups. Thus, in our example, the ratio between individuals erroneously classified as recidivists (FP) and those erroneously classified as non-recidivists (FN) should be equal for Blues and Greens. This criterion is aimed at ensuring that no group will be favored by the system's misclassifications.

With regard to SAPMOC I, this condition is clearly not satisfied, since for Blues the ratio between detrimental errors (FP) and favorable errors (FN) is about four times higher than for Greens, as reported in Table 11.

If is worthwhile noting that some criteria are interdependent, in the sense that if equality under one of them is satisfied, then equality under the other is satisfied as well, For instance, conditional procedure accuracy equality for the positive class is equivalent to the equality of false negative rates and, vice versa, equality of opportunity (i.e., conditional procedure accuracy equality for the negative class) is equivalent to equality of false positive rates.

3.7.6 Preliminary considerations on "fairness" criteria

The analysis of the above criteria is quite puzzling. SAMOC I is apparently unbiased: it focuses on a single feature (having a criminal record) which in both groups is equally correlated with recidivism, and in both groups, it always links having or missing this feature to the same classifications. However, only calibration and false rates parity are satisfied. The different base rates in the two groups prevents other criteria from being satisfied.

In the following, we address this puzzling situation through a general discussion of binary classifications and the ways in which they may affect different groups.

4 Some general facts about binary prediction and group-parity criteria

Here, we provide a general account of the way in which a binary predictive system addresses different groups. Our analysis concerns the domain also explored by Berk et al. (2018) and Kleinberg et al. (2016), as we will show that the achievement of group-parity criteria poses constraints on the statistics of the underlying population, which leads to certain mutual exclusions and possible tradeoffs. Though several sophisticated results can be derived working in this direction, we here highlight the most straightforward ones that can be arrived at using the differentiated thresholding technique we shall propose.

We assume that our system may commit errors. An ideally accurate performance (FN = FP = 0) can never be achieved in practice, this owing to the statistical nature of the relationship between the observed features (and thus the scores based on them) and the outcomes to be predicted.

The confusion matrices for the whole population and for the Blue and Green groups are strictly related since they cover the same individuals. Hence, we may lay out the following set of equalities and inequalities:

$$N^{\mathrm{B}} + N^{\mathrm{G}} = N^{\mathrm{P}} \tag{1}$$

$$\mathrm{TP}^{\mathrm{B}} + \mathrm{TN}^{\mathrm{B}} + \mathrm{FN}^{\mathrm{B}} + \mathrm{FP}^{\mathrm{B}} = N^{\mathrm{B}} \tag{2}$$

$$\mathrm{TP}^{\mathrm{G}} + \mathrm{TN}^{\mathrm{G}} + \mathrm{FN}^{\mathrm{G}} + \mathrm{FP}^{\mathrm{G}} = N^{\mathrm{G}} \tag{3}$$

$$\mathrm{TP}^{\mathrm{B}} + \mathrm{TP}^{\mathrm{G}} = \mathrm{TP}^{\mathrm{P}} \tag{4}$$

$$\mathrm{TN}^{\mathrm{B}} + \mathrm{TN}^{\mathrm{G}} = \mathrm{TN}^{\mathrm{P}} \tag{5}$$

$$\mathrm{FN}^{\mathrm{B}} + \mathrm{FN}^{\mathrm{G}} = \mathrm{FN}^{\mathrm{P}} \tag{6}$$

$$\mathrm{FP}^{\mathrm{B}} + \mathrm{FP}^{\mathrm{G}} = \mathrm{FP}^{\mathrm{P}} \tag{7}$$

$$\mathrm{TP}^{\mathrm{B}}, \mathrm{TN}^{\mathrm{B}}, \mathrm{FN}^{\mathrm{B}}, \mathrm{FP}^{\mathrm{B}} \geq 0 \tag{8}$$

$$\mathrm{TP}^{\mathrm{G}}, \mathrm{TN}^{\mathrm{G}}, \mathrm{FN}^{\mathrm{G}}, \mathrm{FP}^{\mathrm{G}} \geq 0 \tag{9}$$

which define the set of possible individual counts in the three confusion matrices.

As shown in Sect. 5, parity standards are commonly based on a function of the confusion matrix, which we call *focus* and indicate as f. A predictive system, then, provides group parity relative to a focus f when the application of f to the confusion matrix for the Blue group and for the Green group yields the same value, i.e., when

$$f\left(\mathrm{TP}^{\mathrm{B}},\ \mathrm{TN}^{\mathrm{B}}, \mathrm{FN}^{\mathrm{B}}, \mathrm{FP}^{\mathrm{B}}\right) = f\left(\mathrm{TP}^{\mathrm{G}},\ \mathrm{TN}^{\mathrm{G}}, \mathrm{FN}^{\mathrm{G}}, \mathrm{FP}^{\mathrm{G}}\right)$$

Focuses are defined as ratios between two linear combinations (a and b) of the confusion matrix entries, i.e., as

$$\begin{aligned} & f(\mathrm{TP}, \mathrm{TN}, \mathrm{FN}, \mathrm{FP}) \\ &= \frac{a(\mathrm{TP}, \mathrm{TN}, \mathrm{FN}, \mathrm{FP})}{b(\mathrm{TP}, \mathrm{TN}, \mathrm{FN}, \mathrm{FP})} \\ &= \frac{a_{\mathrm{TP}}\mathrm{TP} + a_{\mathrm{TN}}\mathrm{TN} + a_{\mathrm{FN}}\mathrm{FN} + a_{\mathrm{FP}}\mathrm{FP}}{b_{\mathrm{TP}}\mathrm{TP} + b_{\mathrm{TN}}\mathrm{TN} + b_{\mathrm{FN}}\mathrm{FN} + b_{\mathrm{FP}}\mathrm{FP}} \end{aligned} \tag{10}$$

for proper choices of the binary coefficients a_{TP},a_{TN}, a_{FN}, $a_{\mathrm{FP}} \in \{0, 1\}$ and b_{TP},b_{TN}, b_{FN}, $b_{\mathrm{FP}} \in \{0, 1\}$. Among all the possible choices we restrict our attention to those in which $a \neq b$, and the number of non-null coefficients in b is not smaller than the number of non-null coefficients in a. The family $\mathcal{F}$ of the focuses with these features is wide enough to contain most of the fairness criteria proposed in the literature (Kleinberg et al. 2016; Berk et al. 2018).

With this definition, we can prove a simple property: if the two groups are treated equally by the predictive systems from a certain point of view, then they are treated in the same way as the whole population.

Property 1 *For any focus $f \epsilon \mathcal{F}$, if the prediction system is such that if*

$$f\left(\mathrm{TP}^{\mathrm{B}}, \mathrm{TN}^{\mathrm{B}}, \mathrm{FN}^{\mathrm{B}}, \mathrm{FP}^{\mathrm{B}}\right) = f\left(\mathrm{TP}^{\mathrm{G}}, \mathrm{TN}^{\mathrm{G}}, \mathrm{FN}^{\mathrm{G}}, \mathrm{FP}^{\mathrm{G}}\right)$$

then

$$f\left(\mathrm{TP}^{\mathrm{B}}, \mathrm{TN}^{\mathrm{B}}, \mathrm{FN}^{\mathrm{B}}, \mathrm{FP}^{\mathrm{B}}\right) = f\left(\mathrm{TP}^{\mathrm{P}}, \mathrm{TN}^{\mathrm{P}}, \mathrm{FN}^{\mathrm{P}}, \mathrm{FP}^{\mathrm{P}}\right)$$

where $\mathrm{TP}^{\mathrm{P}}, \mathrm{TN}^{\mathrm{P}}, \mathrm{FN}^{\mathrm{P}}, \mathrm{FP}^{\mathrm{P}}$ are the entries in the confusion matrix for the total population.

Property 1 reveals that all parity requirements can be trivially satisfied when all groups have the same statistics as the whole population. When different groups have different statistics—as happens in real cases—satisfying one criterion means failing to satisfy other criteria.

Let us consider, for example, the statistical parity criterion with focus

$$f(\mathrm{TP},\mathrm{TN},\mathrm{FN},\mathrm{FP}) = \frac{\mathrm{TP}+\mathrm{FP}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FN}+\mathrm{FP}} \tag{11}$$

Group-parity requires that the fraction of individuals with a positive prediction is identical in each group. To achieve this result, i.e., to force this invariance, for groups having different base rates, it is necessary to misalign predictions from the probability that the members of the two groups possess the predicted features. For instance, in the SAPMOC example, having both groups with a positive rate (11) of 33% would require considering as non-recidivist 500 of the 1000 Blue individuals having a criminal record (and thus a higher probability of recidivism), while considering all Green individuals in the same condition as recidivists. In groups in which individuals do not share the same input features (predictors) in the same proportions, it is necessary to treat individuals differently in the two groups who share the same features.

Some focuses f' and f'' may be defined so that, if the predictor is fair with respect to the criterion implied by f', it is also fair with respect to the criterion implied by f'', and vice versa. When this does not happen, we will say that the focuses are independent. For independent criteria we have the following property, whose proof is in the appendix.

Property 2. *For any two independent focuses f' and f'', if a prediction system is fair with respect to f' and f'', then its statistical behaviour changes from one group to another only by scaling with respect to the number of individuals, as in the following expressions*:

$$\begin{aligned}
\mathrm{TP}^{\mathrm{B}} &= \eta\mathrm{TP}^{\mathrm{P}}\\
\mathrm{TN}^{\mathrm{B}} &= \eta\mathrm{TN}^{\mathrm{P}}\\
\mathrm{FN}^{\mathrm{B}} &= \eta\mathrm{FN}^{\mathrm{P}}\\
\mathrm{FP}^{\mathrm{B}} &= \eta\mathrm{FP}^{\mathrm{P}}\\
\mathrm{TP}^{\mathrm{G}} &= (1-\eta)\mathrm{TP}^{\mathrm{P}}\\
\mathrm{TN}^{\mathrm{G}} &= (1-\eta)\mathrm{TN}^{\mathrm{P}}\\
\mathrm{FN}^{\mathrm{G}} &= (1-\eta)\mathrm{FN}^{\mathrm{P}}\\
\mathrm{FP}^{\mathrm{G}} &= (1-\eta)\mathrm{FP}^{\mathrm{P}}
\end{aligned} \tag{12}$$

where $\eta = N^{\mathrm{B}}/N^{\mathrm{P}}$.

Note how Property 2 implies that the satisfaction of any two independent criteria entails the satisfaction of any other parity criterion. Yet, to obtain this, the population must be completely homogeneous, and this clearly is not a realistic assumption, nor is it a reasonable requirement to place on real-world prediction systems.

5 SAPMOC II. Scores and thresholds

SAPMOC I worked on the basis of a binary feature, i.e., the presence of a criminal record. To adequately implement the decision model discussed in Sect. 2, we need to slightly complicate the model and assume that the score depends on multiple input features. For instance, we could assume that the score depends on predictors such as the number of previous convictions, age, and psychological profile. Indeed, almost all binary classifiers rely on a score and most of the design effort is usually spent in modelling the possibly complicated function that maps features to scores. For simplicity's sake, here we do not model the connection between input features and scores, but will only assume that the scores are distributed along a scale. The scores can cover any interval (for instance, from 1 to 1000, as in Figs. 2 and 3), but here we assume that the SAPMOC II score is a number in the interval [0, 1] and it is designed so that the likelihood of recidivism increases as the score increases.

5.1 Score densities and SAPMOC II's performance

The relationship between the score and the likelihood of recidivism can be expressed using *densities*. The recidivist density $d_{\mathrm{R}}(s)$ is a function of the score, and its graphic is such that the area beneath the curve between any two values $s' < s''$ is the number of recidivists that are given a score s in the interval $[s', s'']$. The non-recidivist density $d_{\overline{\mathrm{R}}}(s)$ is a function of the score, and its graphic is such that the area beneath the curve between any two values $s' < s''$ is the number of non-recidivists who are given a score s in the interval $[s', s'']$.

Intuitively, densities are histograms with infinitesimal width bars modelling fine-grain details of the distribution of the score across the population. As an example, consider Figure 5, where we represent densities with respect to the SAPMOC II score s of recidivists (red curves) and of non-recidivists (gray curves) for the Blue group (upper pair) and for the Green group (lower pair). In each graph, the number of recidivists/non-recidivists whose scores lie within the interval $[s', s'']$ corresponds to the emphasized area beneath the corresponding section of curve. Clearly, by assuming $s' = 0$ and $s'' = 1$, we are counting all recidivist/non-recidivists, such that Fig. 5 tells us that in the Blue group there are 900 recidivists and 600 non-recidivists (60% base rate),

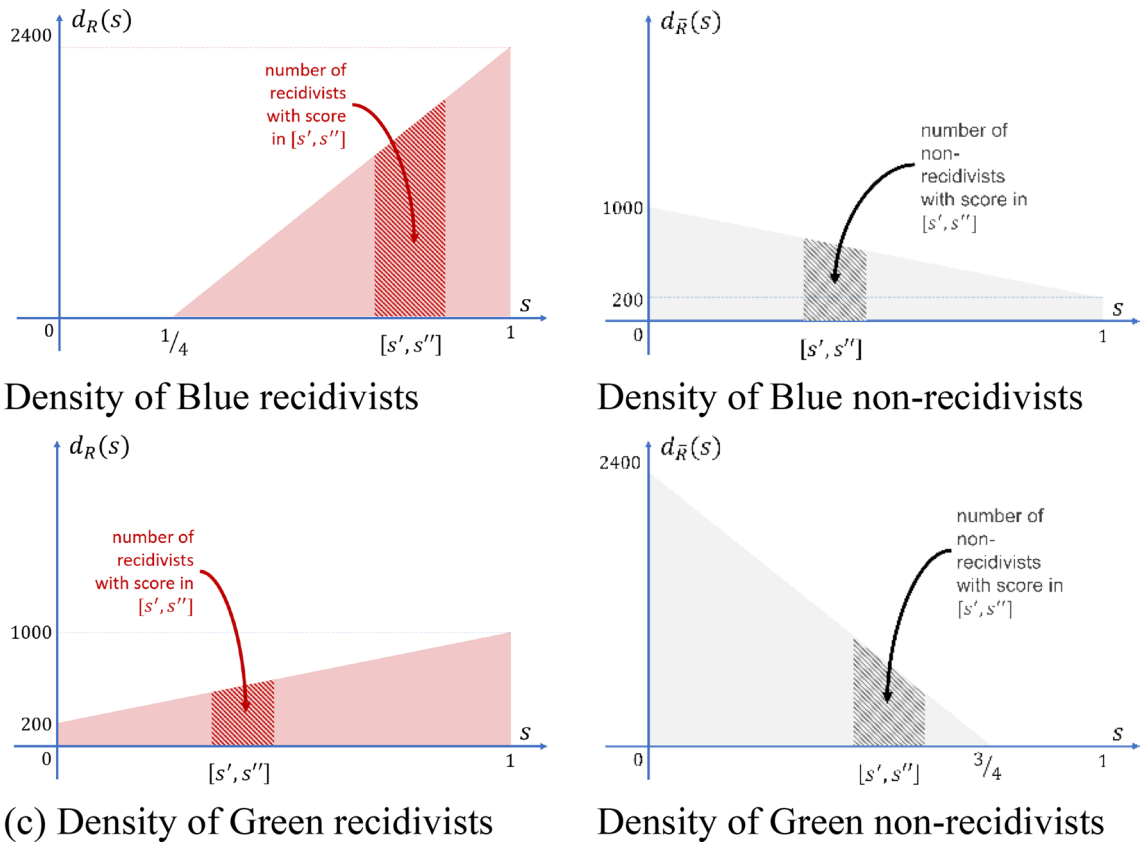

Fig. 5 Densities with respect to the SAPMOC score s of recidivists (red curves (**a**) and (**c**)), and of non-recidivists (cyan curves (**b**) and (**d**)), for the Blue group (upper pair (**a**) and (**b**)), and for the Green group (lower pair (**c**) and (**d**))

while in the Green group we have 600 recidivists and 900 non-recidivists (40% base rate).

The fact that the score is designed to express the likelihood to reoffend reflects the trends of the densities: the density of recidivists has most of the area over large values on the score axis s, while the density of non-recidivists has most of the area over small values on the same axis.

Like most binary decision systems, SAPMOC II matches the score s with a threshold t. Individuals with a score higher than t are predicted to be recidivists, while individuals with a score lower than t are predicted to be non-recidivists. The confusion matrix is defined by the mechanism summarized in Fig. 6, in which the areas corresponding to (the number of) the individuals concerned are highlighted. Errors are committed for all recidivists whose score is lower than t (False Negatives) and for non-recidivists whose score is higher than t (False Positives).

With this insight on SAPMOC II, we recall the implications of Property 1 and consider, for example, the *equal opportunity* criterion, whose focus

$$f(\mathrm{TP}, \mathrm{TN}, \mathrm{FN}, \mathrm{FP}) = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{13}$$

may be read as the probability that an individual is predicted to be a recidivist when he/she will actually reoffend. Assuming that predictions for both groups only depend on the score, equality with respect to this criterion would imply that the individuals in the Blue and Green groups who will reoffend have the same probability of having a score no lower than t. This is a strict constraint, since the predictors' statistics may be very different between groups, as exemplified by the different densities we laid out in Fig. 5.

As a final example, consider the *calibration* criterion, whose focus

$$f(\mathrm{TP}, \mathrm{TN}, \mathrm{FN}, \mathrm{FP}) = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

may be read as the probability that a positive prediction will be true. Again, due to the SAPMOC II mechanism, equality with respect to this criterion implies that the probability that an individual with a score no lower than t be a recidivist must be equal in the Blue group and in the Green one.

To assess the merit and rationale of fairness standards, we need to consider the reasons why a predictive system may deliver different outcomes for different groups.

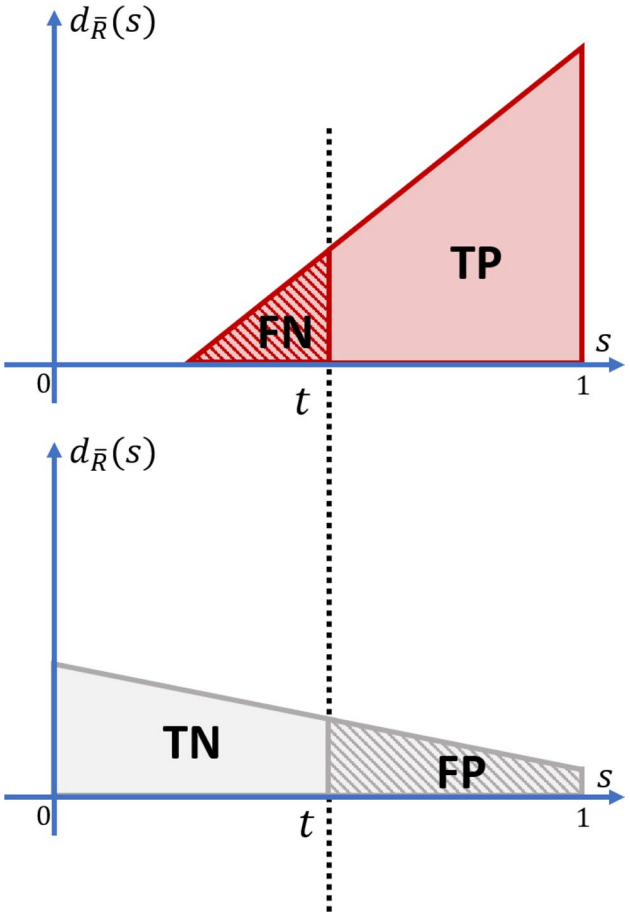

Fig. 6 From densities to confusion matrix by thresholding

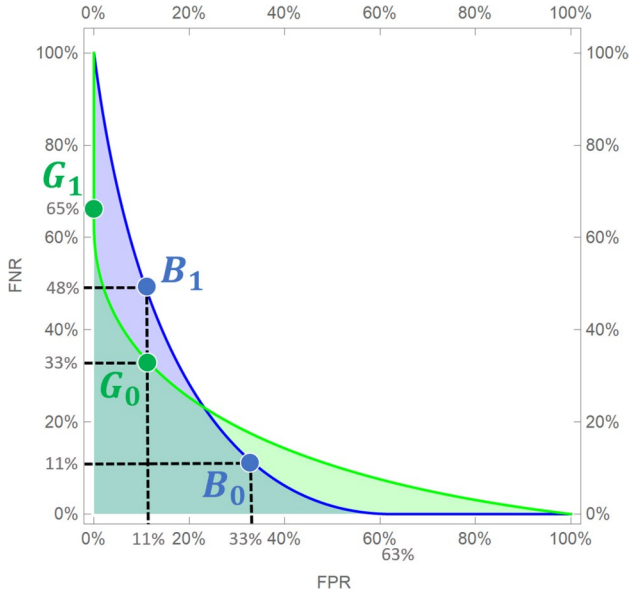

Fig. 7 DET curves for the Blue group and the Green group

Failure to meet a group-parity standard may depend on the training set being biased (for instance, a group may appear to reoffend less, since crimes in that group are less frequently detected). The selection of predictors may also be biased, in the sense that certain predictors may be less correlated with the target in one group than in the other group (e.g., drug abuse, or lack of education, may be more strongly correlated with criminal behavior in one group than in the other). On the other hand, the failure to meet a group parity standard may depend on a different base rate, such that, as in SAPMOC II, the different statistics reflect this ground truth. If the different statistics only depend on a different base rate, rather than on data or predictors being biased, calibration should hold, but the other equality criteria may still be violated.

Different base rates may indeed exist between different groups, e.g., between offenders in different age ranges (young people having a higher propensity to reoffend than older people), between different genders (men having a higher propensity than women), or between different races (as in the COMPAS case). Such differences may depend on a multiplicity of factors, which in some cases (as for the difference between different racial groups) may be connected to social injustices.

5.2 From scores to outcomes via thresholds

A simple and transparent method for implementing policy goals through a predictive system consists in establishing appropriate classification thresholds. In this way, different confusion matrices can be obtained.

Figure 6 clarifies that there is a trade-off between false positives and false negatives. If threshold t for high recidivism is increased, the false negative area (the leftmost part of the recidivist density, indicating the number of individuals wrongly classified as non-recidivists) increases, while the false positive area (the rightmost part of the non-recidivist density, indicating the number of individuals wrongly classified as recidivists) decreases.

To capture such a trade-off and assess how well the decision-making procedure addresses it, we may express errors in prediction as probabilities considering on the one hand the proportion of erroneous predictions relative to non-recidivist, i.e., *false positive rate* $\mathrm{FPR} = \mathrm{FP}/(\mathrm{TN} + \mathrm{FP})$ and on the other hand the proportion of erroneous predictions relative to recidivists, i.e., the *false negative rate* $\mathrm{FNR} = \mathrm{FN}(\mathrm{TP} + \mathrm{FN})$. These two probabilities depend on t, and their pair (FPR, FNR) can be used as coordinates for a point on a plane that changes as t varies.

Letting t assume all possible values, we obtain a set of points in that plane that is commonly indicated as the *Detection Error Trade-off* (DET) curve. Fig. 7, shows the DET curves for the Blue group and for the Green group,

starting from the densities in Fig. 5. The DET curve is fully contained in the unit square and connects the two points (FPR = 0%, FNR = 100%) and (FPR = 100%, FNR = 0%) with a downward trend that represents the trade-off between avoiding false positives and false negatives.

For some intuitive insight on the information provided by DET curves, assume that we start from a threshold $t = 0.5$, which leads to the confusion matrices in Table 5 and to the false rates reported in Table 8. This setting is represented by points B_0 and G_0 in Fig. 7, and reflects the idea that it is equally important to avoid false positives and false negatives. We may think that a false positive rate of 33% in the Blues is too high and that we should therefore reduce it. This can be done by increasing the threshold so that only individuals with very high scores are classified as recidivists. For example, to reduce $\mathrm{FPR}^{\mathrm{B}}$ down to 11% (the original $\mathrm{FPR}^{\mathrm{G}}$), we need to increase the threshold for Blues to $t = 0.77$, thus sliding along the blue DET curve from point B_0 to point B_1. Hence, the decrease in $\mathrm{FPR}^{\mathrm{B}}$ causes an increase in $\mathrm{FNR}^{\mathrm{B}}$ that reaches 48%. In parallel, adopting the same threshold for the Greens causes a movement toward the new point G_1 with $\mathrm{FPR}^{\mathrm{G}} = 0\%$ (no Green individual is erroneously predicted to be a recidivist) and a parallel increase in FNR^{G} to 65%.

This example highlights two phenomena. First, different statistical features (densities) in different groups imply different DET curves and thus different performances of the prediction system. In general, since the lower the error rates, the better, one may assess the quality of the scoring with the area beneath the DET curve (the shaded regions in Fig. 7), a zero area identifying a DET curve that contains the point FPR = FNR = 0% and thus can yield perfect predictions. In our example, the symmetry between the two groups' statistics causes the DET curves to also be symmetric and with an equal underlying area. This means that, in principle, the scoring procedure is equally accurate in the two groups.

Second, despite the equivalent accuracy in the two groups, adopting the same threshold in both leads to different outcomes (in the sense of a different balance between false positives and false negatives) and, more generally, to different confusion matrices. This is coherent with the discussion in Sect. 6 on different criteria for assessing the behavior of predictive systems in groups with potentially different statistical features, and it focuses attention on an internal detail of the prediction mechanism: threshold t.

5.3 Group-dependent thresholding

As noted in Sect. 2, the computation of a score is conceptually distinct from the final decision. In taking the decision, less automated procedures, possibly based on human judgement, may help in adapting statistical assessments to different policy objectives, possibly reflecting social realities.

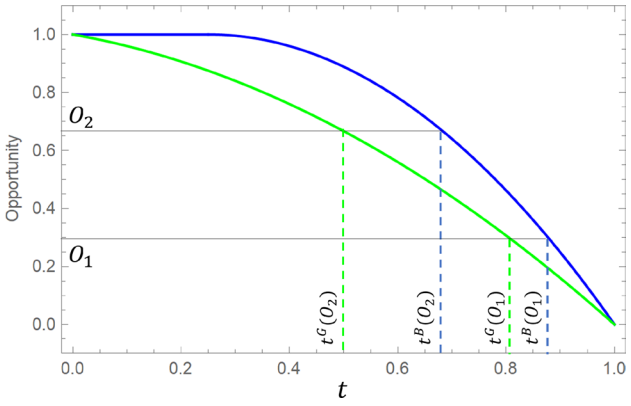

Fig. 8 Opportunity for the Blue and Green groups with score densities as in Fig. 5

The adaptation may consist in basing the classification on different thresholds for different groups.

To explore this direction, assume that we may set two thresholds, t^B and t^G, to give the final prediction for the Blue and Green groups, respectively. In light of the discussion in Sect. 6, we can see that setting different thresholds means decoupling the confusion matrix for the Blue group from the one for the Green group. Though a complete formal treatment of this new degree of freedom is beyond the scope of this paper, it is worthwhile to notice that decoupling such matrices relaxes some constraints to which the matrices are jointly subject in the case of a single threshold. More specifically, since there is no unique threshold, the confusion matrix for the whole population is not defined per se, but is the sum of the confusion matrices for the Blues and for the Greens. Hence (4)–(7), though still valid, are no longer constraints.

This paves the way for satisfying even multiple criteria without making overly unrealistic assumptions about non-modelled uncertainty in classification.

The simple structure of SAPMOC II allows us to exemplify this point. Figure 8 reports the trends of the focus functions for the *equal opportunity* (13) criterion against the value of the threshold t, for the Blue and Green groups. We assume that the densities are those reported in Figure 5.

Clearly, no choice of a unique threshold would be able to ensure group parity with respect to such a criterion, since the two curves do not intersect except at $t = 0$ or t=1, which are unrealistic values.

Yet it is also intuitive that for any level O of the opportunity focus, we may find two values for t^{B} and t^{G} at which the corresponding curves have the same value. This is a quite general property that makes it possible to tune t^{B} and t^{G} independently to satisfy a single group-parity criterion, avoiding the need for the underlying population statistics to satisfy special constraints, as formalized by Property 1.

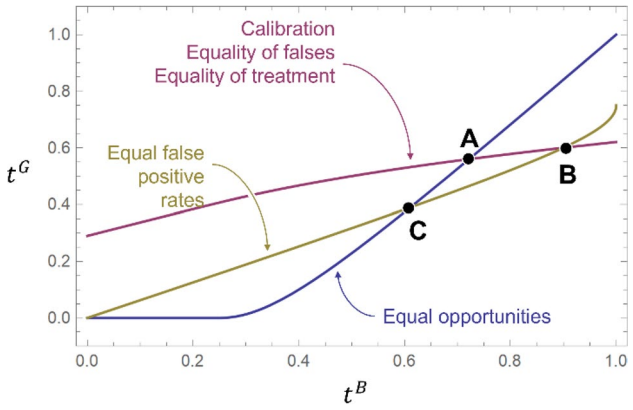

Fig. 9 Satisfaction curves for different criteria with score densities as in Fig. 5

Independent thresholding may even allow two group-parity criteria to be simultaneously satisfied. To see how, consider, for each criterion, a plot like Fig. 8, and for every possible level of a certain focus function, take note of the two thresholds t^B and t^G that are required to satisfy that criterion. The set of (t^B, t^G) pairs defines a *satisfaction curve* in the unit square. Whatever point we consider on such a curve, the criterion is satisfied if we adopt the corresponding thresholds for the two groups.

We may then plot the satisfaction curves for all the criteria of interest on the same plane and obtain something like Fig. 9, which displays the curves for the calibration criterion, the equal opportunity criterion, and for the false negative rate. Every intersection of two of those curves yields a pair of thresholds that simultaneously guarantee group parity with respect to the corresponding criteria. In our example, it is possible to design a prediction system that is both calibrated and yields equal opportunities (point *A* in Fig. 9), or one that is calibrated and results in equal false positive rates (point *B* in Fig. 9), or one that yields equal opportunity and results in equal false-positive rates (point *C* in Fig. 9). Note that this may happen without requiring the complete statistical uniformity that, in the case of a single threshold, would be implied by Property 2.

6 Thresholds and policy goals

In Sects. 5.2 and 5.3 we saw how the outcomes of a binary predictive system can be changed by modifying its thresholds, without interfering with its scoring mechanism.

First, we saw that by raising the threshold for the whole population, false positives diminish and false negatives increase, while the opposite happens by decreasing the threshold. Thus, a threshold setting reflects the comparative importance of avoiding errors in the two classes (i.e., false positives and false negatives) and the corresponding costs. In our example, the cost of a person being erroneously classified as recidivist (and thus, e.g., being denied release on parole) must be compared with the cost of that person being erroneously considered non-recidivist (and thus committing a crime upon release). Note, however, that the cost of the misclassification depends on its legal and social implications and may be assessed differently depending on the different preferences and values. Similar considerations apply in other domains, such as medical diagnosis, where avoiding a false negative is usually much more important than avoiding a false positive.

Second, we saw that by setting different thresholds for different groups, different statistical-equality criteria can be satisfied. In SAPMOC II, using the same threshold for the two groups, we satisfied the calibration criterion, meaning that in both groups an equal likelihood of recidivism gave rise to the same score and therefore to the same classification. As we have shown, the other side of calibration is equality of false rates.

Other group-equality standards can be implemented by setting different thresholds for the two groups. Obviously, this involves differential treatment of individuals sharing the same score, including those having the same values relative to all input features used by the system (except for their group type). The different treatment of individuals sharing the same score may be justified in certain context, either to remedy biases affecting the input data or the selection of features, or according to policy goals that require affirmative action (for a discussion, see Wachter et al. 2021).

As an example of bias in the input data consider, for instance, the case in which a group is subject to more careful controls, so that instances of recidivism are detected to a greater extent (for further discussion on algorithmic biases and policing see, for instance, Chouldechova (2017) and Oswald and Babuta (2019)). In this case, the predictive system will mirror the historical prejudices and inequalities embedded in the input data. Because of the training process, members of the more controlled group may be assigned a higher score than members of other groups, equally likely to reoffend. As an example of a biased selection of features, consider a system predicting that a curriculum will be successfully completed given favorable factors that only apply to a certain group, such as attendance at expensive top schools (for further discussion on algorithmic biases and education see Zeide (2017) and Regan and Jesse (2019)). In both cases, the calibration of the system relative to real outcomes (i.e., the alignment of predictions and probabilities) may be obtained by setting different thresholds.

As an example of policy goals to be achieved, consider the goal of increasing diversity or balancing access to education, types of jobs, or positions. In such cases, the desired

results could be achieved by setting lower thresholds for disadvantaged groups, so that group-parity is achieved, or the distance from such parity is reduced as desired. However, where scores in the two groups are calibrated, i.e., equally correlated with the predicted classifications, introducing different thresholds will entail diminished accuracy of the system (at least for one of the two groups).

Whether the equalization of two groups under certain parity standards—or at least a reduced distance between them—should be aimed at, depends on the purpose of the binary classification as well as on policy goals. Whether a predictive system is made fairer by enforcing such criteria is highly context-dependent. To further qualify this point, consider a system that aims to predict the onset of a disease (e.g., diabetes) so as to take appropriate health measures (e.g., clinical testing and diet recommendations). Assume that the disease is more prevalent in a group in which causal factors positively correlated with that disease (e.g., obesity or some genetic factors) are more frequent. Achieving statistical parity (or another parity standard other than calibration) would require decreasing the number of individuals with a positive prediction in the more affected sub-group or increasing that number in the less affected sub-group.

Similar considerations would apply to a system aimed at detecting socio-economic hardship so as to provide aid. Equalizing under any parity criterion (other than calibration and false rate) would mean considering a greater number of individuals in the wealthier group and/or a lower number of individuals in the poorer group to be in a situation of hardship. Thus, in these two examples, applying group-parity criteria would entail questionable implications: (a) different classifications for individuals (belonging to different subgroups) having the same level of risk; (b) a greater number of erroneous predictions, entailing wrong decisions; and (c) possibly increased differences in health or welfare between the sub-groups.

On the other hand, different thresholds could justifiably be adopted in certain contexts for purposes of affirmative action. This applies especially when a limited number of advantageous positions (e.g., access to certain curricula or jobs) need to be allocated. In such cases, members of the underrepresented groups may be granted a lower threshold for accessing such positions. If the system is calibrated, this should entail a diminished accuracy of its predictions, since individuals more likely to possess the target property (successful completion of the curriculum) are substituted by individuals less likely to have it. However, accuracy may not diminish if the system is not calibrated with regard to changing social realities (in which the disadvantaged groups are more likely to satisfy the prediction in comparison to the instances of them in the training set), so that different thresholds may even ensure more accurate predictions.

In cases pertaining to law enforcement and crime prevention, adopting different thresholds for different groups seems difficult to justify. Thus, calibration is likely to be the key group-fairness criterion. Let us assume that the social and individual costs associated with extending detention for predicted recidivists are deemed excessive because of the way in which detention affects the false positives (those who would not reoffend). For instance, extended detention based on an erroneous prediction of recidivism may reinforce the criminal attitudes of the concerned individuals, make their social reintegration more difficult or adversely affect their families. Such costs can be reduced by raising the threshold for all groups, thereby diminishing positive predictions. This would provide a greater benefit to the group having a larger proportion of predicted recidivists (in comparison to the other group), since raising the threshold would mean that a larger number of members of that group would no longer be predicted to be recidivist. On the other hand, as noted above, raising the threshold would decrease the accuracy of the system: the decrease in mistaken prediction of recidivism would be matched by a larger increase in mistaken prediction of non-recidivism.

Finally, it is possible to improve the fairness of a decision-making process without interfering with classifications, i.e., by intervening on the last step of the decision-making process, namely, the determinations based on such classifications. In the recidivism domain, the socio-legal consequences associated with prediction of recidivism (and in particular of violent recidivism) can be changed, for instance, by substituting extended detention with re-integrative measures and controls compatible with releasing the offender (see Barabas et al 2018). This approach would be comparatively more beneficial to the group with a larger number of predicted recidivists. More generally, we may question the very idea that individuals predicted to have higher probability to engage in future crimes should be targeted with measures that adversely affect them. This idea has been challenged by argueing that such an approach is inherently unfair: given the correlation between social deprivations (poverty, lack of education, unemployment, etc.) and criminality, it selectively harms disadvantaged individuals, while contributing little to the reduction of crime (Harcourt 2008). However, such considerations may not apply to the use of predictive tools for identifying situations of risk that are addressed by supportive measures (van Eijk 2020).

7 Deconstructible AI and awareness in joint human–machine decision

In this section, we argue that the idea we have presented—i.e., aligning decision-making systems with policy goals by explicitly adjusting thresholds—can be generalized to enable finer analyses of decision-making processes and more targeted interventions. To this end, different stages in the

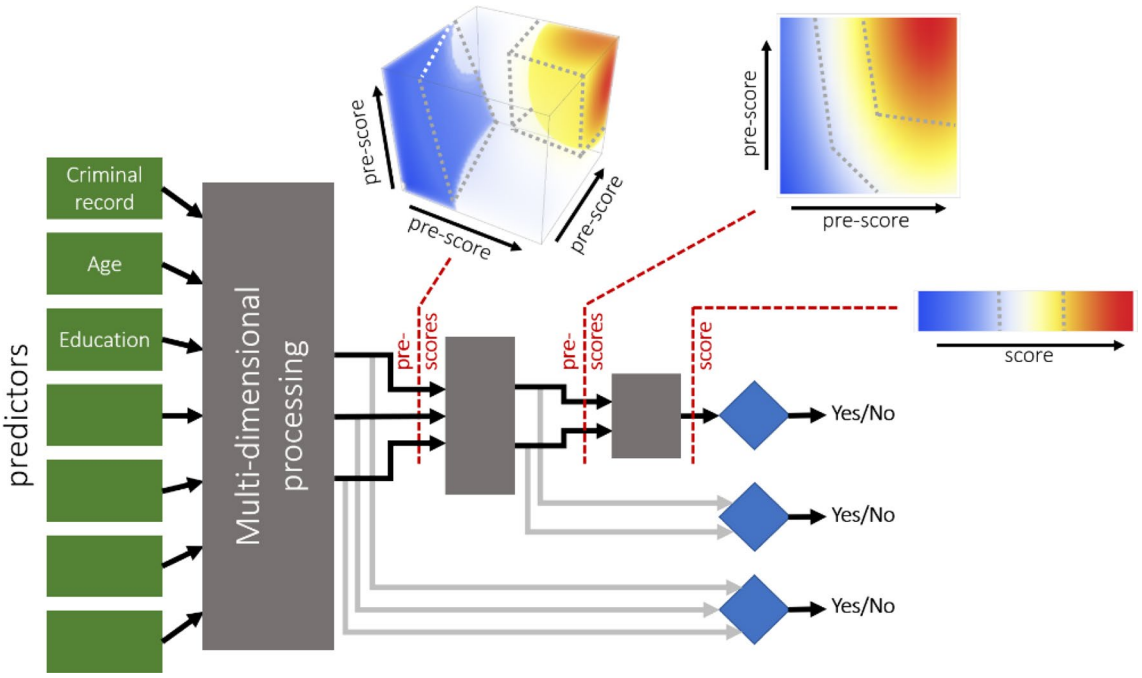

Fig. 10 Progressive exposition of AI internals along with their statistical characterization

automated decision-making need to be distinguished, so that the outcome at each stage can be an object of specific analyses and interventions. This approach may favour the flexible and transparent deployment of predictive systems, their adaptation to social goals and standards of justice, and their critical assessment.

It is indeed true that a binary prediction system, in its most general embodiment, accepts many predictors as inputs, each commonly encoded as a cluster of scalar quantities, and that this system performs potentially complicated calculations that produce a single output bit corresponding to its Yes/No prediction. However, these calculations are often organized in stages (think, for example, of a deep neural-prediction system): a certain number of computations that are performed sequentially, where the results of the previous step are the input for the following one. Usually, the first part of the system's computation are highly dimensional, i.e., they receive a large set of inputs, and send out a large set of outputs. On the contrary, in the final stages the inputs—which we may call pre-scores—are transformed into a smaller number of outputs, working up to the last stage, which computes the single score on which the final prediction/classification is based. This structure is illustrated in Fig. 10, in which we exemplify the last two stages of such a system, i.e., one taking the three intermediate quantities resulting from a potentially multidimensional processing and reducing them to two quantities, and the subsequent one taking these two quantities and producing the final score.

In the examples in Sects. 5.1 to 5.3, the computation of scores in SAPMOC II were viewed as a single process, delivering a score based on the input data. In SAPMOC II, a higher score indicates a greater likelihood of recidivism, and one or more thresholds can be set to produce the binary prediction. By pairing the statistical characterization with thresholds and applying this paring to the total population or to subgroups, we obtain confusion matrices for the prediction system. In turn, confusion matrices give information on the quality of the decision and may be used to verify or enforce group-parity criteria or other desired standards.

Though practical feasibility will depend on the structure of the prediction system, in principle nothing prevents this deconstruction from proceeding further and exposing the last two or even three stages, as we do at the bottom of Fig. 10 by imagining grey paths from the inner stages to the final prediction blocks. In Fig. 10, we indeed assume that the last stages of this process can be separately analysed.

The outcome to be predicted has a relationship with the quantities produced by earlier stages, which in our case are represented by the three- and two-dimensional heat maps[5] at the top of Fig. 10. Note that on multidimensional maps, regions leading to different predictions may have shapes that

[5] The heat maps used here represent densities as defined in Sect. 7, possibly generalized to densities with respect to more than one pre-score, and have to be considered only as an intuitive summary of the elements involved in building the final prediction.

are much more sophisticated than simple thresholding. Such regions give rise to confusion matrices that are defined by greater degrees of freedom, potentially making it possible to address complicated trade-offs between accuracy, fairness, etc., as well as to accommodate predictions subject to political or social constraints. Providing this deconstruction may allow for an informed interaction between the prediction system and its users, the latter being able to match the final decisions with the desired statistical effects on the target population, by playing with the prediction regions.

Though far beyond the scope of this paper, design guidelines might be set out stressing the need for machines whose last stages are easy to interpret or at least to characterize from a statistical point of view., Notice that, though this approach is not equivalent to requiring that a prediction system be explainable per se,[6] it may make such a system more flexible and acceptable in some applicative environments, such as those involving critical judicial or societal problems.

8 Conclusion

In this paper, we discussed group-parity criteria as fairness standards for automated prediction and decision-making. Throughout our inquiry we have used the COMPAS system, complemented by radical simplifications of it (our SAPMOC I and SAPMOC II models), as our running example.

We distinguished three stages in a (partially or totally) automated decision-making process—scoring, classifying, and deciding—and have argued that each stage may require specific interventions to ensure fairness as well as other policy goals.

We then focused on group-parity standards and on their application to assess the fairness of decisions concerning individuals in the justice domain. We introduced the COMPAS system and the debate on whether its use reveals unlawful or unethical discrimination. To exemplify such issues, we presented in detail a simpler system, SAPMOC I, grounding a prediction of recidivism in a single factor (i.e., criminal record). To this system's outcomes we applied multiple group-parity metrics. This analysis has shown how SAPMOC I, like COMPAS, satisfies the calibration criterion: it does so by uniformly treating individuals in different groups having the same probability of recidivism. However, the system fails to implement other group-parity standards. This led us to consider the connection between group-parity standards and the commonsense and philosophical concepts of fairness. We refined our analysis of confusions matrices by examining the connection between the satisfaction of group-parity criteria and base-rates in different groups.

Then we turned to a more complex model, SAPMPOC II, which aggregates multiple predictors in a score along a continuous scale. We observed how the system's correct and erroneous predictions are distributed across the two groups along the ROC curves, and how by shifting thresholds we can change predictions, and consequently modify the sets of individuals who stand to be positively and negatively classified. We observed that there is a trade-off between diminishing/increasing false positives and decreasing/increasing false negatives. We considered how parity along some criteria—or a diminution of distances—can be achieved by adopting different thresholds for different groups.

We then focused on the rationale of modifying thresholds to meet group-parity standards. We observed that the violation of such standards may be related to biases in the data or in the system predictors, or rather to different base rates in the populations. We claimed that modifying thresholds to achieve or move closer to parity standards only makes sense in some contexts, in connection with policy aims. Consequently, we considered that decisional processes supported by predictive systems should allow for human analysis and intervention to adapt them to policy goals as needed.

We hope our analysis may contribute to the current debate on the fairness of predictive systems. The analysis suggests a careful and to some extent skeptical view of group-parity standards. Such standards do not substitute the (debatable, controversial, and fuzzy) notions of fairness in commonsense understanding and in political/philosophical debates. They rather point to all those cases in which a system delivers relevantly different outcomes for different groups. Whether these differences point to unfairness in the decision-making process requires further analysis. Differences between the statistical distribution of the scores assigned to different groups may depend on the fact that the predictive system is not calibrated: it assigns to the members of certain groups scores that under or over-estimate the probability that they possess the predicted property. Alternatively, the system may unduly affect disadvantaged groups when the threshold for ascribing an unfavourable prediction is set too low for everybody, which entails many false positives, mostly affecting the disadvantaged groups.

A system which is calibrated, and whose thresholds are appropriately set, may still violate group-parity criteria to the disadvantage of some groups. Whether having different thresholds for different groups may be desirable, providing a benefit to disadvantaged groups without entailing an unacceptable differential treatment of individuals in other groups, requires ethical and policy considerations that typically pertain to the logic of affirmative action. In any case, the parity criteria we have considered can at most be viewed as proxies for fairness or as clues to possible instances of injustice,

[6] High-dimensional processing is inherently unexplainable except in presence of strong regularities that reduce the number of effective dimensions.

whose ascertainment and remedy requires deeper inquiries. Remedying possible instances of unfairness does not require "optimizing" the concurrent satisfaction of the group-parity criteria proposed in the literature, but rather demands a tailored intervention on the decision-making process, in full awareness of its social function and impacts on individuals.

Appendix

Proof of Property 1 Assume that the total number of individuals in the population is normalized to 1 and consider TP^{P}, TN^{P}, FN^{P}, FP^{P}, TP^{B}, TN^{B}, FN^{B}, FP^{B}, TP^{G}, TN^{G}, FN^{G}, FP^{G}, N^{B}, and N^{G} as free variables.

We know that these free variables must satisfy (1)–(9) in Sect. 4 above, which can be normalized so that the total population is set to $N^{P} = 1$. The proof can be automatized by spanning all $f \epsilon \mathcal{F}$ and, for each of them, using any algebraic manipulation tool (we used Mathematica, Inc. W. R. (2020)) to reduce the following system of equalities and inequalities that should be satisfied simultaneously.

(1) – (9)

$$f\left(\mathrm{TP}^{\mathrm{B}}, \mathrm{TN}^{\mathrm{B}}, \mathrm{FN}^{\mathrm{B}}, \mathrm{FP}^{\mathrm{B}}\right) = f\left(\mathrm{TP}^{\mathrm{G}}, \mathrm{TN}^{\mathrm{G}}, \mathrm{FN}^{\mathrm{G}}, \mathrm{FP}^{\mathrm{G}}\right)$$

$$f\left(\mathrm{TP}^{\mathrm{G}}, \mathrm{TN}^{\mathrm{G}}, \mathrm{FN}^{\mathrm{G}}, \mathrm{FP}^{\mathrm{G}}\right) \neq f\left(\mathrm{TP}^{\mathrm{P}}, \mathrm{TN}^{\mathrm{P}}, \mathrm{FN}^{\mathrm{P}}, \mathrm{FP}^{\mathrm{P}}\right)$$

For each case, the *reduction* proves that the above conditions are incompatible with each other. Hence, if one knows that

$$f\left(\mathrm{TP}^{\mathrm{B}}, \mathrm{TN}^{\mathrm{B}}, \mathrm{FN}^{\mathrm{B}}, \mathrm{FP}^{\mathrm{B}}\right) = f\left(\mathrm{TP}^{\mathrm{G}}, \mathrm{TN}^{\mathrm{G}}, \mathrm{FN}^{\mathrm{G}}, \mathrm{FP}^{\mathrm{G}}\right)$$

then it must also be

$$f\left(\mathrm{TP}^{\mathrm{G}}, \mathrm{TN}^{\mathrm{G}}, \mathrm{FN}^{\mathrm{G}}, \mathrm{FP}^{\mathrm{G}}\right) = f\left(\mathrm{TP}^{\mathrm{P}}, \mathrm{TN}^{\mathrm{P}}, \mathrm{FN}^{\mathrm{P}}, \mathrm{FP}^{\mathrm{P}}\right)$$

Proof of Property 2 First, note that, thanks to Property 1, fairness with respect to f' and f'' implies

$$f'\left(\mathrm{TP}^{\mathrm{B}}, \mathrm{TN}^{\mathrm{B}}, \mathrm{FN}^{\mathrm{B}}, \mathrm{FP}^{\mathrm{B}}\right) = f'\left(\mathrm{TP}^{\mathrm{P}}, \mathrm{TN}^{\mathrm{P}}, \mathrm{FN}^{\mathrm{P}}, \mathrm{FP}^{\mathrm{P}}\right)$$

$$f'\left(\mathrm{TP}^{\mathrm{G}}, \mathrm{TN}^{\mathrm{G}}, \mathrm{FN}^{\mathrm{G}}, \mathrm{FP}^{\mathrm{G}}\right) = f'\left(\mathrm{TP}^{\mathrm{P}}, \mathrm{TN}^{\mathrm{P}}, \mathrm{FN}^{\mathrm{P}}, \mathrm{FP}^{\mathrm{P}}\right)$$

$$f''\left(\mathrm{TP}^{\mathrm{B}}, \mathrm{TN}^{\mathrm{B}}, \mathrm{FN}^{\mathrm{B}}, \mathrm{FP}^{\mathrm{B}}\right) = f''\left(\mathrm{TP}^{\mathrm{P}}, \mathrm{TN}^{\mathrm{P}}, \mathrm{FN}^{\mathrm{P}}, \mathrm{FP}^{\mathrm{P}}\right)$$

$$f''\left(\mathrm{TP}^{\mathrm{G}}, \mathrm{TN}^{\mathrm{G}}, \mathrm{FN}^{\mathrm{G}}, \mathrm{FP}^{\mathrm{G}}\right) = f''\left(\mathrm{TP}^{\mathrm{P}}, \mathrm{TN}^{\mathrm{P}}, \mathrm{FN}^{\mathrm{P}}, \mathrm{FP}^{\mathrm{P}}\right)$$

Assume now that TP^{P}, TN^{P}, FN^{P}, and FP^{P} are fixed, so that $f'_{\mathrm{P}} = f'\left(\mathrm{TP}^{\mathrm{P}}, \mathrm{TN}^{\mathrm{P}}, \mathrm{FN}^{\mathrm{P}}, \mathrm{FP}^{\mathrm{P}}\right)$ and $f''_{\mathrm{P}} = f''\left(\mathrm{TP}^{\mathrm{P}}, \mathrm{TN}^{\mathrm{P}}, \mathrm{FN}^{\mathrm{P}}, \mathrm{FP}^{\mathrm{P}}\right)$ can be considered as constants.

By distinguishing the numerator a' and the denominator b' of f', as well as the numerator a'' and the denominator b'' of f'', we may lay down the following set of linear equalities that must be satisfied simultaneously.

(1) – (7)

$$a'\left(\mathrm{TP}^{\mathrm{B}}, \mathrm{TN}^{\mathrm{B}}, \mathrm{FN}^{\mathrm{B}}, \mathrm{FP}^{\mathrm{B}}\right) - f'_{\mathrm{P}} b'\left(\mathrm{TP}^{\mathrm{B}}, \mathrm{TN}^{\mathrm{B}}, \mathrm{FN}^{\mathrm{B}}, \mathrm{FP}^{\mathrm{B}}\right) = 0$$

$$a'\left(\mathrm{TP}^{\mathrm{G}}, \mathrm{TN}^{\mathrm{G}}, \mathrm{FN}^{\mathrm{G}}, \mathrm{FP}^{\mathrm{G}}\right) - f'_{\mathrm{P}} b'\left(\mathrm{TP}^{\mathrm{G}}, \mathrm{TN}^{\mathrm{G}}, \mathrm{FN}^{\mathrm{G}}, \mathrm{FP}^{\mathrm{G}}\right) = 0$$

$$a''\left(\mathrm{TP}^{\mathrm{B}}, \mathrm{TN}^{\mathrm{B}}, \mathrm{FN}^{\mathrm{B}}, \mathrm{FP}^{\mathrm{B}}\right) - f''_{\mathrm{P}} b''\left(\mathrm{TP}^{\mathrm{B}}, \mathrm{TN}^{\mathrm{B}}, \mathrm{FN}^{\mathrm{B}}, \mathrm{FP}^{\mathrm{B}}\right) = 0$$

$$a''\left(\mathrm{TP}^{\mathrm{G}}, \mathrm{TN}^{\mathrm{G}}, \mathrm{FN}^{\mathrm{G}}, \mathrm{FP}^{\mathrm{G}}\right) - f''_{\mathrm{P}} b''\left(\mathrm{TP}^{\mathrm{G}}, \mathrm{TN}^{\mathrm{G}}, \mathrm{FN}^{\mathrm{G}}, \mathrm{FP}^{\mathrm{G}}\right) = 0$$

by setting appropriate values for the ten free variables TP^{B}, TN^{B}, FN^{B}, FP^{B}, TP^{G}, TN^{G}, FN^{G}, FP^{G}, N^{B}, and N^{G}.

By direct inspection of the corresponding matrix rank, it is easy to ascertain that six out of the seven equalities (1)–(7) are linearly independent. By assumption, the last four linear equalities in the last system of equations above are also linearly independent. From this we get that this system has a unique solution that may be easily verified to be (12).

In such a solution, the entries of the confusion matrix for the Blue and Green groups are proportional to those of the confusion matrix of the whole population. Since any $f \epsilon \mathcal{F}$ is defined as the ratio of linear combinations (10), its value is identical for the Blues and for the Greens.

Funding Open access funding provided by Alma Mater Studiorum - UniversitÃ di Bologna within the CRUI-CARE Agreement.

Acknowledgements Francesca Lagioia and Giovanni Sartor have been supported by the H2020 European Research Council (ERC) Project CompuLaw under the European Union's Horizon 2020 research and innovation programme (Grant Agreement no. 833647) and by the European Union's Justice Programme (2014-2020) Project ADELE: Analytics for DEcision of LEgal cases (Grant Agreement no. 101007420)

References

Agrawal A, Gans J, Goldfarb A (2018) Prediction machines. Harvard Business Review Press, Cambridge

Angwin J, Larson J, Mattu S, Kirchner L (2016) Machine bias: there's software used across the country to predict future criminals and it's biased against blacks. ProPublica, May 23. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Accessed 27 Jul 2021

Barabas C, Dinakar H, Ito J, Virza M, Zittrain J (2018) Interventions over predictions: reframing the ethical debate for actuarial risk assessment. In: FAT 2018 proceedings, p 62–76

Barocas S, Crawford K, Shapiro A, Wallach H (2017) The problem with bias: allocative versus representational harms in machine learning. In: 9th Annual conference of the special interest group for computing, information and society

Barocas S, Hardt M, Narayanan A (2021) Fairness and machine learning. fairmlbook.org

Barocas S, Selbst AD (2016) Big data's disparate impact. Calif Law Rev 104:671. https://doi.org/10.15779/Z38BG31

Berk R, Heidari H, Jabbari S, Kearns M, Roth A (2018) Fairness in criminal justice risk assessments: the state of the art. Sociol Methods Res 50(1):3–44. https://doi.org/10.1177/0049124118782533

Binns, R. (2020). On the apparent conflict between individual and group fairness. In: Proceedings of the 2020 conference on fairness, accountability, and transparency, p 514–524

Brennan T, Dieterich W, Ehret B (2009) Evaluating the predictive validity of the COMPAS risk and needs assessment system. Crim Justice Behav 36(1):21–40. https://doi.org/10.1177/0093854808326545

Chouldechova A (2017) Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. Big Data 5(2):153–163. https://doi.org/10.1089/big.2016.0047

Citron DK, Pasquale F (2014) The scored society: due process for automated predictions. Wash l Rev 89:1

De Vos M (2020) The European court of justice and the march towards substantive equality in European Union anti-discrimination law. Int J Discrim Law 20(1):62–87. https://doi.org/10.1177/1358229120927947

Dieterich W, Mendoza C, Brennan T (2016) COMPAS risk scales: demonstrating accuracy equity and predictive parity. Northpoint Inc 7 (7.4), 1.

Flores AW, Bechtel K, Lowenkamp CT (2016) False positives, false negatives, and false analyses: a rejoinder to machine bias: there's software used across the country to predict future criminals. and it's biased against blacks. Fed. Probation 80, 38

Friedman B, Nissenbaum H (1996) Bias in computer systems. ACM Trans Inf Syst (TOIS) 14(3):330–347. https://doi.org/10.1145/230538.230561

Hajian S, Domingo-Ferrer J (2012) A methodology for direct and indirect discrimination prevention in data mining. IEEE Trans Knowl Data Eng 25(7):1445–1459. https://doi.org/10.1109/TKDE.2012.72

Harcourt BE (2008) Against prediction profiling, policing, and punishing in an actuarial age. University of Chicago Press, Chicago

Hardt M, Price E, Srebro N (2016) Equality of opportunity in supervised learning. arXiv preprint arXiv:1610.02413

Hellman D (2020) Measuring algorithmic fairness. Va Law Rev 106:811

Hildebrandt M (2020) The issue of bias. The framing powers of ML. In: Pelillo M, Scantamburlo T (eds) Machine learning and society: impact, trust, transparency. MIT Press, Cambridge

Inc. W. R. (2020) Mathematica. Version 12.2. Champaign, IL

Joseph M, Kearns M, Morgenstern J, Neel S, Roth A (2016) Rawlsian fairness for machine learning. arXiv preprint arXiv:1610.09559. 1(2)

Kleinberg J, Mullainathan S, Raghavan M (2016) Inherent trade-offs in the fair determination of risk scores. arXiv preprint arXiv:1609.05807

Kusner MJ, Loftus JR, Russell C, Silva R (2017) Counterfactual fairness. arXiv preprint arXiv:1703.06856

Larson J, Mattu S, Kirchner L, Angwin J (2018) How we analyzed the COMPAS recidivism algorithm, ProPublica, May 23. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm. Accessed 27 July 2021

Liptak A (2017) Sent to prison by a software program's secret algorithms, New York Times, May 1. https://www.nytimes.com/2017/05/01/us/politics/sent-to-prison-by-a-software-programs-secret-algorithms.html. Accessed 27 Jul 2021

Mayer-Schönberger V, Ramge T (2018) Reinventing capitalism in the age of big data. Basic Books, New York

O'Neil C (2016) Weapons of math destruction: how big data increases inequality and threatens democracy. Crown, New York

Oswald M, Babuta A (2019) Data analytics and algorithmic bias in policing, Royal United Services Institute for Defence and Security Studies. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/831750/RUSI_Report_-_Algorithms_and_Bias_in_Policing.pdf

Rawls J (2001) Justice as fairness: a restatement. Harvard University Press, Cambridge

Regan PM, Jesse J (2019) Ethical challenges of EdTech, big data and personalized learning: twenty-first century student sorting and tracking. Ethics Inf Technol 21(3):167–179. https://doi.org/10.1007/s10676-018-9492-2

Rescher N (2002) Fairness: theory and practice of distributive justice. Transaction Publishers, Piscataway

Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1(5):206–215. https://doi.org/10.1038/s42256-019-0048-x

Tashea J (2017) Courts are using AI to sentence criminals. That must stop now. Wired, March 17. https://www.wired.com/2017/04/courts-using-ai-sentence-criminals-must-stop-now/. Accessed 27 Jul 2021

van Eijk G (2020) Inclusion and exclusion through risk-based justice: analysing combinations of risk assessment from pretrial detention to release. Br J Criminol 60:1080–1097. https://doi.org/10.1093/bjc/azaa012

Vinuesa R, Hossein Azizpour H, Leite I, Balaam M, Dignum V, Domisch S, Felländer A, Langhans SD, Tegmark M, Fuso Nerini F (2020) The role of artificial intelligence in achieving the Sustainable Development Goals. Nat Commun 11(1):1–10. https://doi.org/10.1038/s41467-019-14108-y

Wachter, S., B. Mittelstadt, and C. Russell (2021) Bias preservation in machine learning: the legality of fairness metrics under EU non-discrimination law. West Va Law Rev 123(3): 735-790

Yong E (2018) A popular algorithm is no better at predicting crimes than random people. *The Atlantic.* January 17. https://www.theatlantic.com/technology/archive/2018/01/equivant-compas-algorithm/550646/. Accessed 27 Jul 2021

Zafar MB, Valera I, Gomez Rodriguez M, Gummadi KP (2017) Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In: Proceedings of the 26th international conference on world wide web, p 1171–1180

Zeide E (2017) The structural consequences of big data-driven education. Big Data 5(2):164–172. https://doi.org/10.1089/big.2016.0061

Žliobaitė I (2017) Measuring discrimination in algorithmic decision making. Data Min Knowl Disc 31(4):1060–1089. https://doi.org/10.1007/s10618-017-0506-1