A Justice-Based Framework for the Analysis of Algorithmic Fairness-Utility Trade-Offs

ANONYMOUS AUTHOR(S)

In prediction-based decision-making systems, different perspectives can be at odds: The short-term business goals of the decision makers are often in conflict with the decision subjects' wish to be treated fairly. Balancing these two perspectives is a question of values. We provide a framework to make these value-laden choices clearly visible. For this, we assume that we are given a trained model and want to find decision rules that balance the interests of the decision maker and the decision subjects. We provide an approach to formalize both perspectives, i.e., to assess the utility of the decision maker and the fairness towards the decision subjects. In both cases, the idea is to elicit values from decision makers and decision subjects that are then turned into something measurable. For the fairness evaluation, we build on the literature on welfare-based fairness and ask what a fair distribution of utility (or welfare) looks like. In this step, we build on well-known theories of distributive justice. This allows us to derive a fairness score that we then compare to the decision maker's utility for many different decision rules. This way, we provide an approach for balancing the utility of the decision maker and the fairness towards the decision subjects for a prediction-based decision-making system.

CCS Concepts: • **Applied computing** → **Law, social and behavioral sciences**; • **Computing methodologies** → Machine learning; • **Social and professional topics** → Socio-technical systems.

Additional Key Words and Phrases: group fairness, distributive justice, utility, welfare, egalitarianism, maximin, prioritarianism, sufficientarianism, Pareto front

ACM Reference Format:

Anonymous Author(s). 2022. A Justice-Based Framework for the Analysis of Algorithmic Fairness-Utility Trade-Offs. In *EAAMO '22: ACM conference on Equity and Access in Algorithms, Mechanisms, and Optimization, October 06–09, 2022, Washington DC, USA.* ACM, New York, NY, USA, 15 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Prediction-based decision-making systems are typically implemented or overseen by the decision makers. They decide what the prediction-based decision-making systems are optimized for and what goals they pursue. However, the increasing use of prediction-based decision-making systems has shown that this can easily lead to disadvantages for marginalized groups (see, e.g., [2, 11, 18, 20, 25, 42]). With no consideration given to fairness, these systems are very unlikely to coincidentally be fair. This reveals that there are at least two conflicting perspectives of what a decision-making process should look like: On the one hand, the decision-making process should pursue the decision maker's goal (e.g., to be as efficient or profitable as possible). On the other hand, the decision-making process should be fair towards the decision subjects, i.e., towards the individuals affected by the decisions. Often, these two goals are in conflict [32].[1] Navigating this trade-off is not easy: It requires making the values of the decision maker and the decision subjects explicit — to the point where they can be put into mathematical formulas. Addressing the perspective of decision subjects, the literature on algorithmic fairness has therefore come up with many different so-called "fairness metrics" [39, 40]: Mathematical formulas that — when they are not fulfilled — may indicate unfairness in the given

[1] Note that they are not always in conflict [15]. Assuming that they are conflicting is in itself a normative assumption. The decision maker and the decision subjects could also pursue goals that are more similar.

system. However, it remains an open question which philosophical theories of justice these fairness metrics align with. One might think that this is not an urgent issue: As long as the metrics seem like reasonable conditions for fairness, why should we not just evaluate them all, independently of whose philosophical theory of justice they correspond to? However, existing research shows that fulfilling three popular fairness metrics at the same time is mathematically impossible in practice [7, 13, 33]. Thus one typically needs to choose one metric over the others — but how do we know which metric we should use to evaluate the fairness of a decision-making system? This depends heavily on the context of the application, so a deep understanding of the social context of the application is required.

It is important to note that the debate about appropriate fairness metrics is not a mathematical debate [47, 53]. The plethora of fairness definitions and the conflicts between them stem from the conflicting theories of fairness that they operationalize and which reflect different values [31]. Thus, it is rather a debate about values [31] and one's beliefs about the world [21]. Recent works [23, 53] have highlighted the need for a deliberative process to explicate these values. Wong [53] argues that the choice of fairness metric(s) is a choice of values and thereby inherently political. Consequently, [53] demands a democratization of this choice.

Hence, what is needed for navigating the trade-off between the perspectives of the decision makers and the decision subjects in practice is a process for eliciting and formalizing their values. This is the goal of this paper: We develop a new approach for implementing moral values in prediction-based decision-making systems. Our approach elicits these values from decision makers and decision subjects in five steps and provides a simple way to adapt the prediction-based decision-making system such that it aligns with the agreed-on values.

The central idea of the approach is to specify one's normative preferences regarding five value-laden questions: (1) How should we assess the benefit / harm that the decision maker derives from the decisions? (2) How should we assess the benefit / harm that the decision subjects derive from the decisions? (3) What groups of people have comparable moral claims to receive the same utility, but probably do not receive the same utility? (4) Consider these groups who have the same claim to benefits / harms from the decisions: Should equality between them be enforced at all costs or can some inequality be tolerated? (5) How strongly should fairness be pursued if it comes into conflict with the utility of the decision-maker?

The rest of the paper is structured as follows: First, we highlight related work in Section 2. In Section 3, we describe the general setting of prediction-based decision-making systems including two conflicting dimensions: the decision makers and the decision subjects. In addition to this, we introduce the notation that we will use throughout this paper. In Section 4, we provide background information on distributive justice and describe a common structure of theories of justice. This will show that theories of justice are conflicting on the level of values. We will then apply this structure to a case study in Section 5. For this, we will have to make value-laden choices. Finally, we discuss the limitations and merits of our approach in Section 6.

2 RELATED WORK

Kearns and Roth [32] describe the importance of considering the different perspectives in the design of decision-making systems: The goal of the decision maker and the fairness of it. Several works (e.g., [16, 24, 37]) have highlighted these conflicts and tried to quantify the trade-off between the two goals. However, these works use specific interpretations of the decision maker's goal (such as accuracy) and fairness (common group fairness metrics). In practice, we cannot assume that these specific formalizations represent the moral values of the decision makers and decision subjects in a given context. As Kearns and Roth [32] highlight, the first step to balancing these two perspectives is therefore to make our values explicit. These values should guide how we formalize the decision maker's goal and fairness.

One approach for operationalizing fairness is to choose between existing group fairness metrics. Saleiro et al. [46] and Makhlouf et al. [36], for example, provide a flow chart to choose between fairness metrics while Loi et al. [35] and Baumann and Heitz [9] provide a fairness principle to morally justify the well-known group fairness metrics. To justify the evaluation of some of these metrics, we have to think through what injustices can occur before decisions are made. Friedler et al. [21] have developed a framework for this, which has been extended by Hertweck et al. [28]. In this process, Hertweck et al. [28] found that enforcing a fairness metric is difficult to justify without considering its consequences. Indeed, Hu and Chen [30] showed that enforcing such metrics can actually harm marginalized groups. As not every group derives the same benefit or harm from the same decision, a line of research has developed that claims what matters for fairness is not the distribution of positive decisions, but of the consequences of those decisions, i.e., the *utility* of the decisions or *welfare*. Weerts et al. [52] have therefore extended the framework by Hertweck et al. [28] to consider the utility of the decisions. Heidari et al. [26] have proposed welfare-based definitions of fairness that take the effects of decisions into account and can be used as learning constraints. Further, [12] have proposed a method to increase the welfare of the worse-off groups according to the Rawlsian leximin principle. Heidari et al. [27] have highlighted that what a fair distribution of utility looks like is influenced by one's claim to utility. In their mapping of the philosophical theory of equality of opportunity to group fairness metrics, they consider the effort to be relevant for the moral claim one has to utility. This argument has been further developed in [35]. They also resolve the apparent conflict between fairness metrics by analyzing fairness at a higher level of abstraction: Appropriate fairness metrics can be directly derived from one's values.

Our approach is in line with these developments as it allows decision makers and decision subjects to derive their own utility-based definitions of fairness from their moral values. Note that these utility-based definitions of fairness are not an alternative to existing group fairness metrics, but an extension of them.[2] We then describe a simple approach that builds on [32] to support the question of how to balance this definition of fairness with the utility of the decision maker.

3 PREDICTION-BASED DECISION-MAKING

This paper is concerned with the fairness of prediction-based decision-making systems that fulfill distributive tasks (e.g., the distribution of loans, job interviews or social benefits). The goal of these systems is to make a decision D based on a set of variables. Predictions are needed because the central variable the decision is based on is not known at the time of decision — we refer to this as the decision-relevant attribute Y. In recruiting, for example, it is unclear whether an applicant will perform well; in medical applications, it is unclear whether a treatment will actually cure the patient. For the purpose of simplification, we assume that D and Y are binary: $D, Y \in \{0, 1\}$. The output of the predictor for a person with the attributes X is then a probability score $p = P(Y = 1|X)$, which is used in the decision-making process. A decision rule r is a function that, for every individual, takes p (and possibly other attributes) as an input and gives a decision as an output, e.g., "give a loan to everyone who has an estimated repayment probability of more than 80%."

In prediction-based decision making systems, the decision maker typically makes many decisions of the same type. Thus, the degree to which the decision maker's goal is achieved can be measured by their expected utility. A rational decision maker would choose the decision rule that maximizes their expected utility. This requires a **first value-laden choice**: The assessment of the decision maker's utility for a given set of decisions. Assuming that the decision maker's expected utility for an individual just depends on the decision and the individuals' probability of being of type $Y = 1$,

[2]This is explained in detail in [3], where it is also shown that existing group fairness metrics — i.e., (conditional) statistical parity, equality of odds, equality of opportunity, FPR parity, sufficiency, predictive parity, and FOR parity — can be derived for specific conditions of our utility-based approach.

the optimal decision rule always takes the form of a single uniform threshold. All individuals with a p above this threshold have a positive expected utility for the decision maker. However, if the fairness of the decisions for the affected individuals should also be considered, the decision maker is required to deviate from their optimal decision rule, as this usually does not satisfy any social desideratum that is unequal to the decision maker's immediate business goal (which is measured by the utility function). This requires the assessment of the decision subjects' utilities for a given set of decisions to specify a morally appropriate definition of fairness — constituting additional value-laden choices, which will be introduced in the following section.

4 THE COMPONENTS OF FAIR DECISION-MAKING

As we explain in the previous section, prediction-based decision making systems are typically driven by the decision maker's perspective. However, in many contexts, the decisions made can have huge impacts on people's lives. Therefore, it is important to also consider another perspective: the fairness of the prediction-based decision-making system towards the decision subjects. This represents a social desideratum, which is highly context-dependent and thus not straightforward to specify. As we are considering systems that distribute something (e.g., loans, social benefits, or jobs), we are building on theories of distributive justice. The goal of theories of distributive justice is to find principles for fair distributions. For this purpose, theories of justice can often be characterized by their answers to the following questions, which represent the **next three value-laden choices**: What is, ultimately, distributed? Between whom is it distributed? How should it be distributed?

4.1 Utility of the decision subjects

What is, ultimately, distributed?

We will refer to what is being distributed, which could be positive in the case of a benefit and negative in the case of a harm, as *utility*. This builds on the line of welfare-based definitions of fairness described in Section 2. Utility can be defined in different ways. We adopt a "desire theory" of utility — utility is what people desire [17]. When we lack information about people's actual desires, we define well-being as what people have reasons to desire — an "informed-desire" approach [17]. Negative utility can be defined as what people desire *not* to have.

Definition 1. *Decision subject utility* is the benefit or harm derived from receiving a certain decision. It can be defined as what people actually desire or what they have reasons to desire.

This general definition can be adapted to different contexts: In some contexts what people desire can be measured in monetary terms. In other contexts, we may measure it on different scales, e.g., as health outcomes.

4.2 Relevant positions

Between whom is it distributed?

Most contemporary theories of justice focus on individuals, understood as bearers of utility, capabilities, or rights [48]. Theories of discrimination, instead, relate to socially salient groups [1]. We can strike a compromise between the two views if we focus on "relevant positions," a concept we adapt from John Rawls [44, §16, pp. 81-86]. In our approach, relevant positions are types of individuals that are representative of salient inequalities in the context to which the question of fairness relates.[3]

[3]Here, we draw from Anonymous [4].

Definition 2. *Relevant positions* are types of individuals that have comparable moral claims to receive the same utility, but probably do not receive the same utility.

In order to identify the relevant positions in a concrete case, we need to focus on two distinct components: 1) what makes it the case that *certain individual types* (classes of people) have roughly the same claims to utility?; 2) what are the most likely sources of inequality? Once these two questions have been answered for a given system, we can define the relevant social positions as the people who have comparable claims to utility, but who we expect not to receive the same utility due to the defined sources of inequalities.

What makes it the case that certain individual types (classes of people) have roughly the same claims to utility? The first question is ethical and must be answered in a context-sensitive manner. For example, there are contexts where individuals who are equal in their needs should be treated equally; in other contexts, it seems morally appropriate to only treat people equally who deserve the same, e.g., because of their actual or potential contribution [38]. In other contexts, yet, everyone should be considered to have an equal claim to what is being distributed [38]. We will refer to this as the *claims differentiator*.

Definition 3. A *claim differentiator* is a factor that makes it that people have equal or different moral claims to utility. This could be factors such as need or contribution. There could also be no claim differentiator, in which case everyone has equal claims to utility.[4]

What are the most likely sources of inequality? This question is, broadly speaking, sociological and asks us to think about different causes of inequality that affect the prediction-based decision-making system.

Definition 4. A *source of inequality* causes inequalities that affect the prediction-based decision-making system. We are considering sources that occur at the group level and that lead to groups being unlikely to receive the same utilities from a prediction-based decision-making system.

One fruitful framework to think about the sources of inequality is the one provided by Friedler et al. [21], which has subsequently been expanded by Hertweck et al. [28]. Friedler et al. [21] propose three layers:

- *Construct Space (CS)*: consists of an individual's characteristics we would like to base decisions on
- *Observed Space (OS)*: consists of the measurements of the characteristics in the construct space that we actually base decisions on, i.e., the features present in a data set
- *Decision Space (DS)*: represents the decisions of the prediction-based decision-making system

Hertweck et al. [28] add one more layer to be considered, at the bottom of the chain of inequality causes:

- *Potential Space (PS)*: represents an individual's innate potential to develop the characteristics in the construct space.

We use the distinction between the spaces to identify the different possible causes of inequalities. Both inequality in the space and in the transition between spaces ("biases") should be considered. Biases occur in the transition between the observed space and decision space ("direct discrimination"), between the construct space and observed space ("measurement bias") and between the potential space and construct space ("life's bias"[5]). If we have reasons to believe that such unjust biases occur, we can consider those to be sources of inequalities. For example, consider fairness in

[4]The claim differentiator corresponds to the *justifier* described by Loi et al. [35].
[5]The framework differentiates between just and unjust life's bias. We would only consider unjust life's bias as a relevant source of inequality as just life's bias is not something we would want to correct for. For a more detailed discussion of the differentiation between just and unjust life's bias, see [28]

hiring. We can identify sex and race as two significant causes of inequality in the transition from the potential space to the construct space, meaning that people with the same innate potential end up developing different realized abilities because of their life circumstances ("life's bias"). The use of this four-level heuristics can lead us to discovering new sources of inequality.

Note that Weerts et al. [52] have further expanded this framework by adding one more layer at the top:

- *Utility Space (US)*: represents the consequences of the decisions in the decision space and how beneficial or harmful they are for the decision subjects. This is referred to as the "utility" of decisions.

While unjust inequalities can occur between the decision space and the utility space, this is not a source of inequality we are looking for in this step. Instead, this inequality can be represented through the previous step (Section 4.1) in which the utility of the decision subjects is assessed. This assessment allows us to ascribe different utilities to individuals who receive the same decisions.

4.3 Relation to inequality

How should it be distributed?

After having discussed which groups have equal moral claims to the utility derived from the decisions, we have to think about whether we can tolerate inequalities in some cases. One may say that inequalities are always unacceptable and that equality has to be achieved at all costs. However, this might result in leveling down: Assume a situation in which the utilities derived for the groups are unequal, but in order to equalize them, the utility of all groups has to be lowered. In that case, one might prefer the original unequal utility distribution, from which all groups profit. This is a well-known issue with existing group fairness metrics (see, e.g., [10, 15, 30, 52]). To avoid this, we can allow for some inequalities, e.g., if they are beneficial to the worst-off group. Therefore, we have to define our relation to inequality. This can be described as a *pattern of justice*.

Definition 5. A *pattern of justice* describes a distribution of utility and thereby how inequality should be dealt with.

Popular patterns of justice are:

- Egalitarianism [5]: Equality is valued above all else. The utilities derived by the groups should be as equal as possible.
- Maximin [44, 45]: Inequalities are tolerated if they profit the worst-off group. The goal is to maximize the utility derived by the group that is worst-off.
- Prioritarianism [29]: Inequalities are tolerated if they increase the aggregated utility. The goal is to maximize the aggregated utility derived by all groups, where the utility of the worst-off is given a higher weight. Maximin is the extreme version of this as an infinite weight is given to the worst-off group.
- Sufficientarianism [50]: Inequalities are tolerated as long as all groups achieve a minimum level of utility. The goal is to bring all groups above a certain level of utility; if inequalities occur above this minimum level of utility, they are accepted.

4.4 Popular theories of justice

We can view many popular theories of justice through the lens of these three components. Egalitarian notions of fairness, for example, demand that people are equal in some regard [5]. Equality of opportunity is the view that individuals should have equal opportunities in life [6]. This is different from strict egalitarianism, which demands equality in

outcomes [34]. They differ in their answer to *what* (component 1) should be distributed equally: opportunities or outcomes (e.g., resources). Luck egalitarianism is a form of egalitarianism that favors equality between individuals, but deems inequalities just if said inequalities are not due to bad luck, but if the individuals are responsible for them [14, 19]. Luck egalitarianism thus uses responsibility as a claim differentiator.

5 CREDIT LENDING EXAMPLE

Let us now see how the perspectives of the decision maker (Section 3) and the decision subjects (Section 4) can be balanced in practice. For this, we will use an example that is commonly used in the algorithmic fairness literature: *lending*. Imagine you work at a bank and your job is to decide which loan applications should be accepted ($D = 1$) and which ones should be rejected ($D = 0$). You have built a machine learning model to help you with the task. It predicts how likely a person is to repay the loan they applied for based on their loan application. The goal is to give those people a loan who will repay it ($Y = 1$). You trained your model with data from previous positive lending decisions, so based on the applications of people you have given a loan to in the past and for whom you thus know whether they repaid it or not. Now you have to define a decision rule. If you want to balance the goal of being fair with the profitability of the credit approval system, you have to specify how you want to measure the bank's utility, i.e., profits, and how you want to measure fairness. To illustrate this example, we are using the preprocessed version of the UCI German credit dataset [22].[6]

5.1 Utility of the decision maker

We start by specifying the perspective of the decision maker. Assuming that the bank is interested in profits, we have to assess how much profit is derived from each decision. For a loan applicant i with a repayment probability of p_i asking for a loan of size s_i with an interest rate of z_i, the bank's expected utility is $E(u_{DM,i}) = p_i \cdot z_i \cdot s_i - (1 - p_i) \cdot s_i$. A rational decision maker would grant a loan to all individuals with a positive expected utility (i.e., $D = 1$ if $E(u_{DM,i}) > 0$). To make this assessment easier, we assume that the bank always demands 10% interest. In this case, a repaid loan gives the bank a profit of 10% of the loan size while the cost of a defaulted loan is equivalent to the loan size. Rejected loan applications are cost-neutral as we assume that the cost of reviewing applications is almost 0.[7] The expected utility of the decision maker for individual i can thus be described as $E(u_{DM,i}) = p_i \cdot 0.1 \cdot s_i - (1 - p_i) \cdot s_i$. Loans are granted to individuals with a positive expected utility:

$$
\begin{aligned}
E(u_{DM,i}) &> 0 \\
p_i \cdot 0.1 \cdot s_i - (1 - p_i) \cdot s_i &> 0 \\
p_i \cdot 0.1 \cdot s_i - s_i + p_i \cdot s_i &> 0 \\
(0.1 + 1) \cdot p_i \cdot s_i - s_i &> 0 \\
1.1 \cdot p_i \cdot s_i &> s_i \\
p_i &> \frac{s_i}{1.1 \cdot s_i} \\
p_i &> \frac{1}{1.1}
\end{aligned}
$$

[6] In order not to compromise the mutually anonymous reviewing process, we cannot yet make the code for this publicly available, but will do so if the paper is accepted.

[7] We need to make assumptions when working with hypothetical cases, but in real-world settings, stakeholders will know what applies to their case. The fact that we need to make such assumptions should be considered a feature, not a limitation. It is often said that fairness is contextual. The fact that we need information about the context to assign utility values shows that our approach is highly sensitive to aspects of the context.

Thus, the bank would apply a decision rule that takes the form of a single uniform threshold and give a loan to all individuals with $p_i > \frac{1}{1.1}$.

5.2 Utility of the decision subjects

Next, we turn to the evaluation of how fair a given decision-making system is towards the decision subjects. For this, we consider the components of fair decision-making described in Section 4.

We first have to answer the question of how to assess the utility that decision subjects derive from the decision-making process. For this, we can use the concept of the utility space introduced by Weerts et al. [52], which we briefly described in Section 4. In the case of lending, *loans* are distributed. We could therefore measure the utility in terms of resources available to the individuals. In that case, all people who are given a loan of the same size derive equal utility from it. However, individuals do not profit equally from being granted a loan. If they cannot repay the loan and end up defaulting, this is harmful for their future chances of receiving a loan. Thus, we decide that to consider the utility of a decision, it is insufficient to only consider the decision itself. Instead, we also have to consider the situation that this puts the individual in. For this, we have to consider whether the individual is actually able to repay the loan. Additionally, we may consider factors such as the loan size (e.g., defaulting on a smaller loan may be less harmful than defaulting on a bigger loan) or any other measurable attributes (e.g., individuals of a marginalized group might profit more from receiving a loan than individuals of a group that is better-off in many aspects of life). To keep this example simple, we will base the assessment of the utility on the decision D and the individual's repayment ability Y. We then have to assign utility values to all combinations of the attribute that we think the utility depends on. In our case, D and Y are binary variables, so there are four combinations that we can assign utilities to. As it is difficult to quantify how much benefit or harm an individual derives in these situations, we will simplify this analysis to assign a value between -10 and +10.

$D = 1$ *and* $Y = 1$. This asks for the utility of an individual who is granted a loan and repays it. Clearly, the individual derives a benefit from this: They receive the loan they applied for and can use it as planned. We therefore assign the maximum utility of +10.

$D = 1$ *and* $Y = 0$. This asks for the utility of an individual who is granted a loan and defaults. As stated above, the individual derives a harm from this: They receive the loan they applied for, but end up in debt as they cannot repay it. We assign a utility of -5.

$D = 0$ *and* $Y = 1$. This asks for the utility of an individual who is not granted a loan even though they would have been able to repay it. Their situation does not change much compared to their current situation. They have to invest additional time to apply for another loan, but assuming that there are other banks who will approve their loan application, this is only a small harm. We therefore assign a utility of -1.

$D = 0$ *and* $Y = 0$. This asks for the utility of an individual who is not granted a loan and would not have been able to repay it. Their situation does not change much compared to their current situation and given that they would not be able to repay their loan, they do not miss an opportunity by not being granted the loan. We therefore consider this combination to be neutral and assign a utility of 0.

5.3 Relevant positions

We now have to define the relevant positions to compare.

For this, we have to identify the claim differentiator, i.e., answer the question "What makes it the case that *certain individual types* (classes of people) have roughly the same claims to utility?" In our example, we can say that people who can repay their loan deserve more utility than people who cannot repay their loan. While those who cannot repay their loan do not deserve to be punished (i.e., they should not receive a negative utility), they do not have a moral claim to profiting from the decisions. Therefore, we will compare the utility of people who repay their loan ($Y = 1$) and who thus have an equal moral claim to utility.

Next, let us turn to the sources of inequalities, so to the question "What are the most likely sources of inequality?" To answer this question, we use the framework of different spaces described in Section 4. In our example, we are limited to the variables in our dataset. We assume that groups defined by the sex attribute[8] have unjustly unequal chances in life (e.g., because women are less likely to be considered for promotions and are more likely to be the main care-taker of their children), leading to different repayment abilities. We therefore suppose that we are dealing with sex as a cause of inequality in the transition from the potential space to the construct space (unjust life's bias).

It follows that the relevant social positions are: people who repay their loans ($Y = 1$) who differ in their sex. We therefore demand equality in the expected utility of women and men who are able to repay their loan.

5.4 Relation to inequality

Inequalities in the utilities of men and women might be seen as an indicator of unfairness. However, enforcing equality can come at a cost to the already disadvantaged group [30]. Instead of enforcing equality, we can instead try to maximize the utility of the worst-off group. This is what we will aim for in our example. Though, it should be noted that reasonable individuals can disagree about what level of inequality can be tolerated. When we compare different decision rules, we can analyze how well they do with respect to our declared fairness goal ("maximize the utility of the worst-off group"). For this, we calculate the expected utilities of all groups for each decision rule and compare the lowest expected utility.

5.5 Trade-off decision

Through the previous steps, we arrive at our definition of a fairness score, which can help us to evaluate the fairness of a decision rule.[9] While we cannot say that a system that scores high in the defined metric is fair beyond this rather narrow definition, we can take low scores as warning signs of unfairness as this means that the utility of the worst-off group is low. In order to see the conflict between fairness and the utility of the decision maker, we plot the decision maker's utility against the fairness score for various decision rules [32].

However, the options are fairly limited if there is only one threshold that applies to every individual. It is easier to, e.g., equalize the expected utility between women and men if the thresholds are allowed to be different for the groups. In this case, every group has their own threshold, so if, e.g., a woman applies for a loan, her score p has to be above the threshold specified for women. This is what we call *group-specific thresholds*. We will evaluate many such group-specific thresholds with respect to the utility of the decision maker and the fairness score they produce.

For this, we test n thresholds for both women and men and combine them in every possible way. We then plot the Pareto front (blue line) of the resulting n^2 threshold combinations as seen in Figure 1.[10] The Pareto front allows us to compare optimal threshold combinations, i.e., decision rules for which an improvement for one dimension is only possible if the other dimension is worsened. How fair we want to be (i.e., which point on the Pareto front is most

[8] Even though sex is not binary, it is represented as a binary variable in this dataset.
[9] The mathematical side of how exactly to derive a fairness score from these fairness components is described in full detail in [3].
[10] In principle, the number of thresholds that can be used for each group is infinite. In practice, we may plot the Pareto front for a very large number of thresholds combinations. Here, we used 101 thresholds for each group, resulting in 10 201 threshold combinations.

desirable), is a question of values (assuming that fairness conflicts with the goals of the decision maker): It represents the **fifth value-laden choice** of our framework. Figure 1 helps us with this choice by making the trade-off more explicit.

As a start, we can look at the two extreme points: the one that maximizes the utility of the decision maker (point (0)) and the one that maximizes the fairness score (point (10)). As can be seen in Figure 2, maximizing the utility of the decision maker results in a low utility for women ($A = 0$). Compared to that, the utility of women is at its maximum for the highest fairness score.

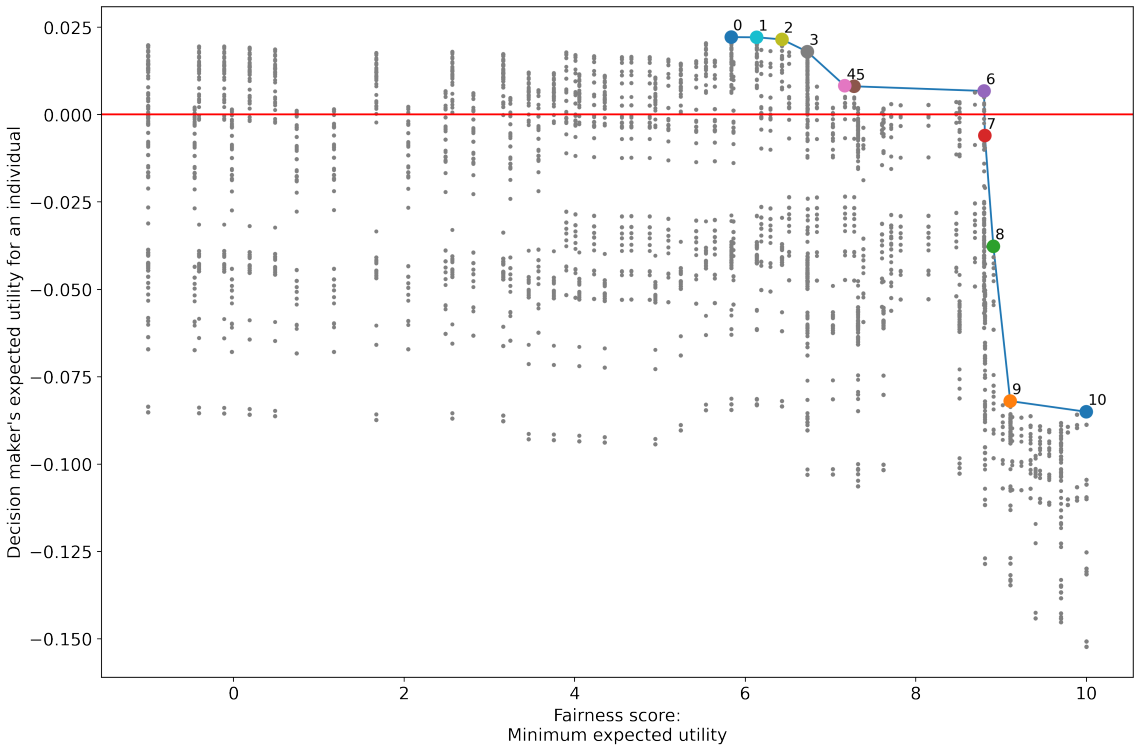

Fig. 1. The Pareto front where the small, gray points are Pareto-dominated by the larger, colored points.

Figures 1 and 2 also show the other points on the Pareto front and the corresponding utilities for women and men. As we can see, some points on the Pareto front would lead to a negative expected utility for the decision maker (points below the red line in Figure 1). Clearly, the bank would not choose such decision rules as they would sooner or later go out of business. Among the other points, one could argue that point (6) represents a good trade-off: It achieves a high fairness score with a high expected utility for both men and women while still being profitable for the bank. From this point on, one has to sacrifice a lot of the fairness score in order to gain a little in the utility of the decision maker (points (4) and (5)), so one could argue that this gain in the utility of the decision maker is too costly in terms of fairness. For similar reasons, point (3) appears to be another justifiable choice.

It is important to note that reasonable people can disagree about the values we specified in this example. From the assessment of utility (for the decision maker and decision subjects), to the choice of relevant positions and a pattern

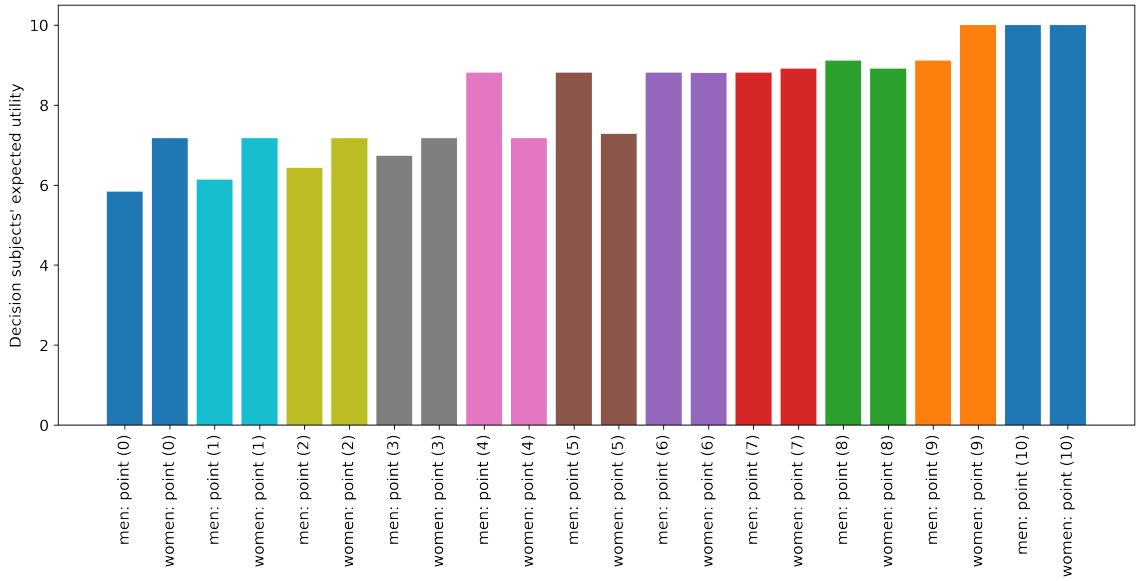

Fig. 2. The utilities of women and men resulting from the decision rules on the Pareto front.

of justice, to the choice of a trade-off: These are all value-laden choices. The goal of our framework is to make these value-laden choices explicit and to find decision rules that align with these values.

6 DISCUSSION

Recall that the design of a prediction-based decision-making system requires us to explicate our moral values. In this paper, we proposed a framework for eliciting these moral values in five steps. These five steps ask decision makers and decision subjects to clarify their normative preferences: They have to specify how to assess the utility of the decision makers and of the decision subjects, the relevant positions, the relation to inequality, and to what degree fairness should be enforced if it is in conflict with the decision maker's utility. Previous attempts to guide stakeholders in choosing appropriate fairness metrics have taken on the form of explicit rules, such as in [36, 46]. Such rules, however, assume a limited set of fairness definitions between which stakeholders can choose. What we provide is instead a method of constructing ad-hoc fairness metrics that reflect the values decided on by the stakeholders. This addresses some limitations of the issues that come with choosing between a narrow set of fairness definitions: Choosing maximin as one's preferred pattern of justice, for example, allows for inequalities if they profit the worst-off group. Moreover, the utility can be defined for each group independently and can depend on every variable we have data on.

So what does this mean in practice? As we have seen, fairness metrics implement *values*. Thus, in practice, it makes sense to first specify one's answers to the five value choices we presented in our framework. Then, one can use the corresponding metric for evaluating and improving the model.[11] We can choose a decision rule that aligns with our

[11]Note that this is a simplification: The metric used to *evaluate* the fairness of a model might be different from the metric we *enforce*. We can use metrics as indicators of unfairness, but due to in-group differences, enforcing this same metric might lead to within-group unfairness. Therefore, it might make sense to choose different metrics for these two purposes.

values using a Pareto front, which compares the decision maker's utility and fairness of different Pareto-optimal decision rules.

6.1 Limitations

In this paper, we only consider (group-specific) thresholds and thereby limit the different types of decision rules one could possibly evaluate. However, there are also other types of decision rules that might be relevant to maximize the decision maker utility while at the same time considering group fairness: For example, Baumann et al. [8] show that (depending on the groups' score distributions and the decision maker's utility function) it can be optimal for the decision maker to apply an upper-bound threshold for one of the groups (i.e., miss out on the most promising individuals of one group in order to select the best individuals in all other groups) if satisfying the group fairness metrics predictive parity or FOR parity is a strict requirement. In addition to this, it has been shown that randomized thresholds are optimal to satisfy more than one group fairness metric at the same time (Hardt et al. [24] show that this is the case for the metrics equality of opportunity and FPR parity — which is called equalized odds or separation — and Baumann et al. [8] show that this is the case for the metrics predictive parity and FOR parity — which amounts to sufficiency).[12] Our approach can easily be extended with these (or other) types of decision rules. Instead of other decision rules, we could also consider other types of bias mitigation techniques such as pre-processing or in-processing.[13] This would lead to additional points in the Pareto plot, which could also be evaluated in terms of the two dimensions considered (the perspectives of the decision maker and the affected individuals). Note that we do not claim that our threshold-based Pareto front leads to the best possible trade-off between the two dimensions. It is possible that, e.g., training a new model with a fairness constraint leads to a better trade-off. The threshold-based Pareto front is merely meant to support the discussion about the values embedded in the trade-off choice.

Our approach also seems to assume that the goal of the decision maker is necessarily in conflict with fairness. Cooper and Abrams [15] warn against this framing of trade-offs between the two perspectives.[14] They argue that this framing does not consider the possibility that fairness and decision maker utility can sometimes go hand in hand. Therefore, we want to highlight that depending on what values are expressed in the utility function of the decision maker and the fairness score, their optimization can actually be mutually beneficial. In that case, the Pareto front might just consist of a single decision rule.

While the approach we presented is very flexible, this flexibility comes with its own limitations: It is difficult to find a utility function that one thinks sufficiently represents the complexity of the decision-making process and its consequences. Moreover, quantifying values such as well-being or freedom is obviously difficult [49]. The same goes for relevant positions and choosing a pattern of justice: Precisely because reasonable people can disagree about which choices are most appropriate, it is difficult to choose just one.

This is perhaps why current group fairness metrics are so tempting: They do not require us to think of a complicated utility function. However, we must not delude ourselves: Not specifying a utility function does not mean that we remain agnostic about what an appropriate utility function is — we simply choose it indirectly. We argue that it is preferable to make these value-laden choices explicit in the design process. Indeed, current group fairness metrics may work well as simplifications of more complex utility functions. Nonetheless, it is important to recognize the values embedded in these metrics.

[12] See Verma and Rubin [51] and Barocas et al. [7] for a detailed explanation of the mentioned group fairness metrics.
[13] We point to Pessach and Shmueli [43] for an overview of different bias mitigation techniques.
[14] They specifically consider accuracy as the decision maker's goal, but their argument extends to a more general goal of maximum utility.

Finally, while our framework is compatible with many theories of distributive justice, it is not compatible with theories that do not follow the patterns of justice described in Section 4. This is, for example, the case for Nozick's entitlement theory [41].

6.2 Future work

This paper describes an approach that is already oriented towards practitioners. However, more work will be required to truly use this approach in practice. Decision makers and decision subjects are unlikely to exhaust all degrees of freedom of our framework (e.g., very complex utility functions and advanced patterns of justice). Rather, such details might deter first-time users from using this framework at all. It is therefore important to distill the fundamental elements of this framework and simplify it enough to make it easily usable while still keeping it flexible enough to allow for more complex scenarios. We are currently working on a web application that will allow decision makers and decision subjects to apply a simplified version of this framework to their own example. For this, we will have to find out how many and which patterns of justice users find helpful and how complex they want the utility function to be. This will have to be tested in practice.

REFERENCES

[1] Andrew Altman. 2020. Discrimination. In *The Stanford Encyclopedia of Philosophy* (Winter 2020 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.

[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias. *ProPublica* (2016). https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[3] Anonymous. 2022. Distributive Justice as the Foundational Premise of Fair ML: Unification, Extension, and Interpretation of Group Fairness Metrics. (2022). Unpublished manuscript.

[4] Anonymous. 2022. Representative Individuals. (2022). Unpublished manuscript.

[5] Richard Arneson. 2013. Egalitarianism. In *The Stanford Encyclopedia of Philosophy* (Summer 2013 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.

[6] Richard Arneson. 2015. Equality of Opportunity. In *The Stanford Encyclopedia of Philosophy* (Summer 2015 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.

[7] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2020. Fairness and Machine Learning. http://fairmlbook.org Incomplete Working Draft.

[8] Joachim Baumann, Aniko Hannak, and Christoph Heitz. 2022. Enforcing Group Fairness in Algorithmic Decision Making: Utility Maximization Under Sufficiency. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (forthcoming)*. Association for Computing Machinery, New York, NY, USA.

[9] Joachim Baumann and Christoph Heitz. 2022. Group Fairness in Prediction-Based Decision Making: From Moral Assessment to Implementation. In *2022 9th Swiss Conference on Data Science (forthcoming)*.

[10] Reuben Binns. 2018. Fairness in Machine Learning: Lessons from Political Philosophy. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, New York, NY, USA, 149–159. http://proceedings.mlr.press/v81/binns18a.html

[11] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR, 77–91.

[12] Violet Xinying Chen and JN Hooker. 2022. Combining leximax fairness and efficiency in a mathematical programming model. *European Journal of Operational Research* 299, 1 (2022), 235–248.

[13] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.

[14] Gerald Allan Cohen. 1989. On the Currency of Egalitarian Justice. *Ethics* 99, 4 (1989), 906–944.

[15] A. Feder Cooper and Ellen Abrams. 2021. Emergent Unfairness in Algorithmic Fairness-Accuracy Trade-Off Research. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (Virtual Event, USA) *(AIES '21)*. Association for Computing Machinery, New York, NY, USA, 46–54. https://doi.org/10.1145/3461702.3462519

[16] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*. 797–806.

[17] Roger Crisp. 2021. Well-Being. In *The Stanford Encyclopedia of Philosophy* (Winter 2021 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.

[18] Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters* (2018). https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

[19] Ronald Dworkin. 1981. What is Equality? Part 2: Equality of Resources. *Philosophy and Public Affairs* 10, 4 (1981), 283–345.

[20] Virginia Eubanks. 2018. *Automating inequality: How high-tech tools profile, police, and punish the poor.* St. Martin's Press.

[21] Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2016. On the (im)possibility of fairness. *arXiv preprint arXiv:1609.07236* (2016).

[22] Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. 2019. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*. 329–338.

[23] Maryam Amir Haeri, Kathrin Hartmann, Jürgen Sirsch, Georg Wenzelburger, and Katharina A Zweig. 2022. Promises and Pitfalls of Algorithm Use by State Authorities. *Philosophy & Technology* 35, 2 (2022), 1–31.

[24] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. *arXiv preprint arXiv:1610.02413* (2016).

[25] Elisa Harlan and Oliver Schnuck. 2021. Objective or biased: On the questionable use of Artificial Intelligence for job applications. *Bayerischer Rundfunk (BR)* (2021). https://interaktiv.br.de/ki-bewerbung/en/

[26] Hoda Heidari, Claudio Ferrari, Krishna Gummadi, and Andreas Krause. 2018. Fairness behind a veil of ignorance: A welfare analysis for automated decision making. *Advances in Neural Information Processing Systems* 31 (2018).

[27] Hoda Heidari, Michele Loi, Krishna P Gummadi, and Andreas Krause. 2019. A moral framework for understanding fair ML through economic models of equality of opportunity. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 181–190.

[28] Corinna Hertweck, Christoph Heitz, and Michele Loi. 2021. On the Moral Justification of Statistical Parity. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 747–757. https://doi.org/10.1145/3442188.3445936

[29] Nils Holtug. 2017. Prioritarianism. In *Oxford Research Encyclopedia of Politics*.

[30] Lily Hu and Yiling Chen. 2020. Fair classification and social welfare. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 535–545.

[31] Abigail Z Jacobs and Hanna Wallach. 2021. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 375–385.

[32] Michael Kearns and Aaron Roth. 2019. *The ethical algorithm: The science of socially aware algorithm design.* Oxford University Press.

[33] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).

[34] Julian Lamont and Christi Favor. 2017. Distributive Justice. In *The Stanford Encyclopedia of Philosophy* (Winter 2017 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.

[35] Michele Loi, Anders Herlitz, and Hoda Heidari. 2021. Fair Equality of Chances. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (Virtual Event, USA) *(AIES '21)*. Association for Computing Machinery, New York, NY, USA, 756–756. Available at SSRN: https://ssrn.com/abstract=3450300.

[36] Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. 2021. On the Applicability of Machine Learning Fairness Notions. *SIGKDD Explor. Newsl.* 23, 1 (may 2021), 14–23. https://doi.org/10.1145/3468507.3468511

[37] Aditya Krishna Menon and Robert C Williamson. 2018. The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency*. PMLR, 107–118.

[38] David Miller. 1999. *Principles of social justice.* Harvard University Press.

[39] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application* 8 (2021), 141–163.

[40] Arvind Narayanan. 2018. Translation tutorial: 21 fairness definitions and their politics. In *Conference on Fairness, Accountability and Transparency*.

[41] Robert Nozick. 1974. *Anarchy, state, and utopia.* Vol. 5038. new york: Basic Books.

[42] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453.

[43] Dana Pessach and Erez Shmueli. 2022. A Review on Fairness in Machine Learning. *ACM Comput. Surv.* 55, 3, Article 51 (feb 2022), 44 pages. https://doi.org/10.1145/3494672

[44] John Rawls. 1999. *A Theory of Justice* (2 ed.). Harvard University Press, Cambridge, Massachussets.

[45] John Rawls. 2001. *Justice as fairness: A restatement.* Harvard University Press.

[46] Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T Rodolfa, and Rayid Ghani. 2018. Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577* (2018).

[47] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*. 59–68.

[48] Amartya Sen. 1980. Equality of what? *The Tanner lecture on human values* 1 (1980), 197–220.

[49] Amartya Sen. 1985. The Standard of Living. *The Tanner lecture on human values* (1985). https://tannerlectures.utah.edu/_resources/documents/a-to-z/s/sen86.pdf

[50] Liam Shields. 2020. Sufficientarianism. *Philosophy Compass* 15, 11 (2020), e12704. https://doi.org/10.1111/phc3.12704

[51] Sahil Verma and Julia Rubin. 2018. Fairness Definitions Explained. In *Proceedings of the International Workshop on Software Fairness* (Gothenburg, Sweden) *(FairWare '18)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3194770.3194776

[52] Hilde Weerts, Lambèr Royakkers, and Mykola Pechenizkiy. 2022. Does the End Justify the Means? On the Moral Justification of Fairness-Aware Machine Learning. *arXiv preprint arXiv:2202.08536* (2022).

[53] Pak-Hang Wong. 2020. Democratizing algorithmic fairness. *Philosophy & Technology* 33, 2 (2020), 225–244.