Distributive Justice as the Foundational Premise of Fair ML: Unification, Extension, and Interpretation of Group Fairness Metrics

ANONYMOUS AUTHOR(S)*

Group fairness metrics are an established way of assessing the fairness of prediction-based decision-making systems. However, these metrics are still insufficiently linked to philosophical theories, and their moral meaning is often unclear. In this paper, we propose a comprehensive framework for group fairness metrics, which links them to more theories of distributive justice. The different group fairness metrics differ in their choices about how to measure the benefit or harm of a decision for the affected individuals, and what moral claims to benefits are assumed. Our unifying framework reveals the normative choices associated with standard group fairness metrics and allows an interpretation of their moral substance. In addition, this broader view provides a structure for the expansion of standard fairness metrics that we find in the literature. This expansion allows addressing several criticisms of standard group fairness metrics, specifically: (1) they are parity-based, i.e., they demand some form of equality between groups, which may sometimes be detrimental to marginalized groups; (2) they only compare decisions across groups but not the resulting consequences for these groups; and (3) the full breadth of the distributive justice literature is not sufficiently represented.

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, fairness metrics, distributive justice, consequential decision-making, machine learning

ACM Reference Format:

Anonymous Author(s). 2023. Distributive Justice as the Foundational Premise of Fair ML: Unification, Extension, and Interpretation of Group Fairness Metrics. In *2023 ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '23), October 30 – November 1, 2023, Boston, US.* ACM, New York, NY, USA, 41 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Supervised machine learning (ML) is increasingly used for prediction-based decision-making in various consequential applications, such as credit lending, school admission, and recruitment. Research has shown that the use of algorithms for decision-making can reinforce existing biases or introduce new ones [6, 39]. Consequently, fairness has emerged as an important desideratum for automated decision-making. As many cases have shown, considering fairness explicitly is crucial in order to avoid disadvantages towards marginalized groups (see, e.g., [2, 12, 16, 24, 52, 58]).

Different measures have emerged in the algorithmic fairness literature for assessing unfairness in decision-making systems, many of which are in the category of so-called group fairness criteria[1]. The concept of group fairness stands in contrast to approaches focusing on individuals, such as individual fairness [19, 64], or counterfactual fairness [43]. This paper focuses on group fairness metrics.

[1]Readers unfamiliar with group fairness may refer to [50], [65], and [5, Chapter 3] for an overview of the topic, and to Appendix A for a brief introduction of the most-discussed group fairness criteria.

Manuscript submitted to ACM

Heidari et al. [27] provide a unifying framework for these criteria. However, they only consider standard fairness criteria that demand equality between different socio-demographic groups, i.e., that are based on an egalitarian notion of distributive justice [11]. They do not discuss non-egalitarian fairness criteria, which – as we will see in Section 2.2 – can be relevant for the assessment of fairness. Kuppler et al. [42] find that "apparently, the fair machine learning literature has not taken full advantage of the rich and longstanding literature on distributive justice" [42, p. 17]. Our paper addresses this gap by building on extensions of standard group fairness criteria and linking them to the distributive justice literature, considering both egalitarian and non-egalitarian concepts. We propose a generalized framework for assessing the fairness of decision systems, drawing on the concept of distributive justice. Based on this, we offer a generalized definition of group fairness, which includes the known group fairness criteria but significantly extends the space of group fairness.

While decision systems are usually designed to optimize a certain goal for a decision maker, they also produce some benefit or harm for the affected individuals. On a societal scale, the repetitive application of the decision system leads to a *distribution of benefit/harm* among different social groups. We study the question of group fairness by building on theories of distributive justice, which are concerned with the question of when such a distribution can be called just. Our suggested framework consists of the following four components:

(1) **Utility of the decision subjects:** Defines how to measure the amount of benefit/harm for decision subjects.
(2) **Relevant groups:** Defines the social groups to be compared with respect to how much utility they receive.
(3) **Claim differentiator:** Defines the features which justify inequalities in the distribution of utility between individuals.
(4) **Pattern of justice:** Defines what constitutes a just distribution.

All four components represent normative choices about what constitutes justice or fairness and are built on existing work. The four components as such are thus not novel but rather an established part of the literature on fairness metrics. The key novelty of our paper is the combination of these existing components into a comprehensive framework for group fairness metrics. This is important for the following reasons:

- **Unification:** We show that the most popular group fairness metrics can be interpreted as instantiations of our framework. Thus, the framework provides a unification of established group fairness metrics, interpreting them as different applications of a common general principle of distributive justice.
- **Extension:** Our framework is built on a generalized definition of group fairness, which establishes the general structure of group fairness criteria and also suggests ways to diverge from established criteria. Therefore, the framework can be used to construct new criteria that are adapted to the context of the application.
- **Interpretation:** Each component in our framework is linked to particular aspects of the moral assessment of a decision-making system. When we interpret established group fairness criteria as special cases of our framework, we can thus explicate the assumptions that are implicitly embedded in these group fairness criteria. Thereby, we provide new insights into established group fairness criteria and make it easier to evaluate whether a fairness criterion is morally appropriate for a given context.

The paper is structured as follows: We first present existing literature on group fairness in Section 2. Specifically, we will discuss the limitations of standard group fairness metrics and how existing work has expanded on these standard metrics. In Section 3, we present our comprehensive framework for group fairness. We focus on the mathematical formalization of different aspects of the distributive justice literature while keeping the review of the philosophical foundations short. More details about the philosophical background can be found in the companion paper [3], which

we provide in Appendix C. Section 4 then demonstrates that standard group fairness metrics are special cases of our group fairness framework. Next, in Section 5, we showcase the extensions of our framework compared to existing approaches using an example from the medical domain. Finally, we discuss the implications of our framework and possible future work in Section 6.

2 RELATED WORK

Our work focuses on group fairness criteria. The most popular group fairness criteria have been developed in the context of binary classification problems and are derived from the confusion matrix: Conditional probabilities such as the true positive rates or the positive predictive values are compared across groups. We refer to these as "standard group fairness criteria" (see Appendix A). In this section, we first take a look at the limitations of the standard group fairness criteria and then discuss how they have been expanded in other works.

2.1 Limitations of standard group fairness criteria

There does not seem to be a clear consensus on what group fairness is and different terms have been used in the literature to describe the concept. To frame our understanding of the literature's current view on (standard) group fairness criteria, we refer to the following definitions:

(1) "Group fairness ensures some form of statistical parity (e.g. between positive outcomes, or errors) for members of different protected groups (e.g. gender or race)" [11, p. 514] (based on [19]'s definition of statistical parity)
(2) "Different statistical fairness criteria all equalize some group-dependent statistical quantity across groups defined by the different settings of [the sensitive attribute] A" [5, Chapter 3].

These definitions show the following three common properties of standard group fairness metrics: (1) they consider multiple groups, (2) they compare averages over groups, and (3) they demand parity between these (what is referred to as *egalitarianism*).[2] Standard group fairness metrics suffer from several limitations:

The "leveling down objection". As shown by [32], enforcing equality can yield worse results for *all* groups. This so-called "leveling down objection" is often brought forward to challenge egalitarianism in philosophical literature [17, 53]: In a case in which equality requires us to worsen the outcomes for everyone, should we really demand equality or should we rather tolerate some inequalities?[3] As criticized by [14] and [66], standard definitions of group fairness lack this differentiation as they always minimize inequality.

Focus on decisions instead of consequences. As pointed out by [28] and [66], standard fairness criteria like statistical parity or equality of opportunity focus on an equal distribution of favorable *decisions* and not on the *consequences* of these decisions. They assume "that a 'positive classification' output is an equally valuable outcome for everyone" as pointed out in [21, p. 491]. Similarly, [10] notes that these criteria "[assume] a uniform valuation of decision outcomes across different populations" [10, p. 6], and highlights that this assumption does not always hold. This makes it difficult to use standard group fairness criteria for a moral assessment of unfairness. Moreover, parity-based criteria do not allow for unequal treatment, even if this may be desirable from a social justice perspective in certain cases [37].

[2]Note that these definitions of group fairness also fall into the category of "oblivious measures" [23, p. 3] and "fairness definitions from data alone" [49, p. 149], i.e., measures that only require access to the data of the decision-making system. This stands in contrast to alternative concepts of fairness that "incorporate additional context" [49, p. 154] (such as individual fairness [19] or causal definitions of fairness [43]), which we do not consider here.
[3]One may argue that in a case where equality is not met, one should opt for, e.g., the collection of better data instead of worsening a group's utility. However, it cannot be ruled out that some form of de-biasing would still be necessary and could then worsen a group's utility. Societal inequalities, for example, persist even in "better" data, so there is no guarantee that equality can be achieved while keeping the same level of utility for all groups.

Limited set of fairness definitions. Standard group fairness metrics differ with respect to their underlying moral values [33, 59]. As they are mathematically incompatible, one has to choose one over the others [5, 13, 40, 41, 67]. None of the standard group fairness criteria proposed in the algorithmic fairness literature might be morally appropriate in a given context (see the example in Section 5).

2.2 Extension of standard group fairness criteria

New group fairness metrics have been suggested to overcome the limitations of standard group fairness metrics. Several works have taken a utility-based view of fairness to overcome the issue that standard fairness metrics do not consider the mapping of decisions to a benefit/harm for decision subjects.[4] Finocchiaro et al. [21] point out that "utilitarianism and normative economics have been extensively used in mechanism design to motivate using utility functions as a synonym for social welfare" and suggest that machine learning could build on this through, for example, "individual and group-level utilities" [21, p. 491]. Individual utilities have been used to define "fairness behind a veil of ignorance" [26]. Several works conceive group-based notions of fairness that are centered on utility [9, 31, 32]. [9] show that enforcing fairness criteria may harm marginalized groups if the wrong utility values are assumed. Similar to our findings, they also point out that standard group fairness metrics map to utility-based group fairness metrics. However, they do not move beyond parity-based notions of fairness.

[31] adapts the concept of envy-freeness to fair machine learning by "requir[ing] that individuals in group G [do] not prefer the classification given to individuals in group $\hat{G}$ (and not just the classification that would be given to them if the classifier for group $\hat{G}$ were used for them)" [31, p. 2]. They show how standard group fairness criteria can be mapped to this interpretation of group-level envy-freeness. While this creates some connection between envy-freeness and our framework, we will explain in Section 6.2 why group-level envy-freeness does not fall neatly into our framework.

Most of the discussed works so far have taken a parity-based approach, which [11, 42] connect to (luck) egalitarianism. [32] already mentions that one may want to maximize the utility of the marginalized group to overcome the "leveling down" objection against parity-based metrics. Indeed, the distributive justice literature offers many more distribution patterns than egalitarianism [42]. Some expansions of group fairness have looked at these other patterns. [47] and [18] do not attempt to equalize harm across groups but to minimize the harm of the group with the highest error rate (referred to as minimax). The diametrically opposed maximin principle, which was popularized by John Rawls [54, 55], maximizes the benefit or utility of the worst-off group. [22] describes the maximin principle as being useful in high-stakes decision-making where risk aversion is appropriate. These works take important steps in considering other distributions of utility than parity-based ones. However, they still only look at one specific pattern and do not discuss how this fits into a general framework of group fairness.

So while several works expand on standard group fairness criteria, none of them provides a comprehensive framework that integrates different theories of distributive justice. The goal of this paper is to propose such a framework.

3 A FRAMEWORK FOR FAIRNESS EVALUATIONS BASED ON DISTRIBUTIVE JUSTICE

We consider a decision-making system that takes binary decisions D on decision subjects (DS) of a given population P, based on a decision rule r. The decision rule assigns each individual $i \in P$ a binary decision $d_i \in \{0, 1\}$, which depends on an unknown but decision-relevant binary random variable Y, by applying r to some input data. The decision rule could, for example, be an automated rule that takes decisions based on predictions of Y via a predicted score $S \in [0, 1]$

[4] Note that some of this literature uses the term "welfare" instead of "utility", which can be traced back to the different fields these works intersect with.

for every individual, derived from an ML model, or the decisions could be made by humans. We assume that at least two socially salient groups are defined, denoted by different values a for the sensitive attribute A. In the following, we first introduce the four components of our framework for group fairness. Then, in Section 3.5, we provide a generalized definition of group fairness, which encompasses all components of the proposed framework.

3.1 Modeling consequences: Utility of the decision subjects

For modeling the consequences of decisions for decision subjects, we use a utility function u_{DS} which, in our binary context, may depend on both the decision d_i and the value y_i of Y. u_{DS} is positive in the case of a benefit, and negative in the case of a harm. In the simplest case, we ignore individual differences in the utility function. Then, the utility $u_{DS,i}$ of a decision subject i with $Y = y_i$ subjected to a decision $D = d_i$ is given by:

$$u_{DS,i} = w_{11} \cdot d_i \cdot y_i + w_{10} \cdot d_i \cdot (1 - y_i) + w_{01} \cdot (1 - d_i) \cdot y_i + w_{00} \cdot (1 - d_i) \cdot (1 - y_i), \tag{1}$$

where w_{dy} denote the four different utility values that might be realized for the four combinations of the random variables Y and D, leading to the utility matrix $W = (w_{00}, w_{01}, w_{10}, w_{11})$. [5] The utility $u_{DS,i}$ is a realization of a random variable U_{DS}. For assessing the fairness of a decision-making system, we are interested in *systematic* differences between groups. We follow the standard group fairness assumption that such differences correspond to different expectation values $E(U_{DS})$ for different groups in A [5].

3.2 Defining groups: Relevant groups

Group fairness is concerned with socially salient groups (e.g., defined by gender or race) as this is what theories of discrimination focus on [1]. We refer to these groups as *relevant groups*, denoted by A, and expect them to at least have a *weak causal influence* on the prediction or outcome (or both). This means we can plausibly expect group membership in the relevant groups to be a (direct or indirect) cause of inequalities. In [3], a philosophical argument is provided for this definition.

3.3 Defining subgroups: Claim differentiator

Comparing the relevant groups as such might not always be morally appropriate. For example, equality of opportunity [23] only considers individuals with $Y = 1$. This might be considered morally appropriate if the moral claim for a positive decision depends on y_i. In our framework, we allow for a so-called *claim differentiator*, represented by a feature J which differentiates individuals with different claims to the utility. Different claims may be justified, e.g., by differences in deservingness, need, or merit. All individuals with the same value $J = j$ are considered to have the same claim to utility. [6] Consequently, comparing relevant groups a may be conditioned on subgroups with an equal moral claim (hence equal value j): Instead of $E(U_{DS}|A = a)$, we compare $E(U_{DS}|J = j, A = a)$. Note that not all possible values j might be considered relevant from a fairness perspective.

[5] More complex utility functions can be used, up to a fully individualized utility function. A simple extension would also take A into account and define the utility matrix for each group separately, i.e., using utility weights of all possible outcomes that depend on the group membership ($w_{dy} \not\perp a$). In philosophy and economics, the work of Amartya Sen explains why resources do not always convert into the same capabilities (options to be and do) [61, pp. 21-23], which would suggest such an extension.

[6] A similar idea is found in [8, 29, 45].

3.4 Just distribution: Pattern of justice

The claim differentiator J defines which individuals have equal moral claims to the utility distributed by the decision process. One might argue that this calls for equal utility. However, the literature on distributive justice shows that this is not necessarily the case. Our approach thus offers additional normative choices, which we refer to as *patterns of justice*. For each of them, we will briefly explain their normative view of what constitutes justice and formulate a *fairness constraint*, representing a mathematical formalization of a pattern of justice, which can either be satisfied or not. For simplicity, we restrict ourselves to the case of two relevant groups $A = \{0, 1\}$ even though our framework generalizes to more groups.

In the following, we introduce only a few patterns of justice (representing fairness principles for the distribution of benefits) that are widely discussed in the philosophical literature. However, our utility-based definition of group fairness should in no way be seen as limited to these patterns.

3.4.1 Egalitarianism. Egalitarianism – as the name suggests – demands equality [4]. However, egalitarianism as a broad concept does not specify *what* should be equalized. This is the subject of the *equality of what* debate initiated by [60]. One could, e.g., aim to equalize the opportunities (equality of opportunity) or outcomes (equality of outcomes). In our approach, we consider utility as the quantity that has to be equalized.

Fairness criterion. The egalitarian fairness criterion is satisfied if the expected utility is equal for the relevant groups conditioned on the claim differentiator:

$$E(U_{DS}|J = j, A = 0) = E(U_{DS}|J = j, A = 1) \tag{2}$$

3.4.2 Maximin. Maximin describes the principle that among a set of possible distributions, the one that maximizes the expected utility of the group that is worst-off should be chosen [44]. In contrast to egalitarianism, inequalities are thus tolerated if the worst-off group benefits from them. This has been defended by Rawls in the form of the "difference principle" [54, 55].

Fairness criterion. A decision rule r' satisfies the maximin fairness criterion if there is no other possible rule r that would lead to a greater expected utility of the worst-off group $E(U_{DS})^{worst-off}$.

$$E(U_{DS})^{worst-off}(r') \geq max_{r \in R}\left(E(U_{DS})^{worst-off}(r)\right) \tag{3}$$

3.4.3 Prioritarianism. Prioritarianism describes the principle that among a set of possible distributions, the one that maximizes the weighted sum of utilities across all people should be chosen, where the utility of the worst-off group is given a higher weight [30]. Thus, the normative goal is to maximize $\tilde{U}_{DS} = k \cdot E(U_{DS})^{worst-off} + E(U_{DS})^{better-off}$, with $k > 1$. [7]

Fairness criterion. A decision rule r' satisfies the prioritarian fairness criterion if

$$\tilde{U}_{DS}(r') \geq max_{r \in R}\left(\tilde{U}_{DS}(r)\right), \tag{4}$$

3.4.4 Sufficientarianism. Sufficientarianism [62] describes the principle that there is a minimum threshold of utility that should be reached by everyone in expectation. Inequalities between relevant groups are acceptable as long as all groups achieve a minimum level of utility in expectation.

[7] The maximin principle can be seen as the extreme version of this as an infinite weight is given to the worst-off relevant groups).

Table 1. Utility matrix representation of metrics used for standard group fairness criteria. Gray patches depict unused DS utility weights due to the claim differentiator $J = j$. The DS utility weights (w_{dy}) are represented as

	Y=0	Y=1
D=0	w_{00}	w_{01}
D=1	w_{10}	w_{11}

U_{DS}	J	j	**Metric**
0 0 / 1 1	$\emptyset$	-	Acceptance rate
0 / 1	Y	$\{1\}$	True positive rate
0 / 1	Y	$\{0\}$	False positive rate
0 1	D	$\{1\}$	Positive predictive value
0 1	D	$\{0\}$	False omission rate

Fairness criterion. The sufficientarian fairness criterion is satisfied if all groups' expected utilities are above a given threshold t:

$$\forall a \in A \;\; E(U_{DS}|J = j, A = a) \geq t \tag{5}$$

3.5 Generalized definition of group fairness

Instead of seeing group fairness as demanding equality between socio-demographic groups with respect to a statistical quantity, we propose the following generalized definition:

Definition 3.1 (Group fairness). *Group fairness* is the just distribution of utility among groups, as defined by the specification of a utility function, relevant groups, a claim differentiator, and a pattern of justice. *Group fairness criteria* specify when group fairness is satisfied by a decision-making system.

We will show that the standard group fairness criteria are special cases of this definition of group fairness with different utility functions and claim differentiators. However, all of them are based on the pattern of egalitarianism. The proposed generalization allows for arbitrary utility matrices, yielding the possibility to compare consequences rather than decisions, and additional patterns of justice, as suggested by the relevant philosophical literature.

This extension of group fairness criteria alleviates some of the criticisms of currently popular group fairness criteria as we will show in Section 6.

4 RELATION TO STANDARD GROUP FAIRNESS CRITERIA

Standard group fairness criteria derived from the confusion matrix are special cases of the group fairness framework that we propose. They follow the egalitarian pattern of justice and correspond to specific decision subject utility functions (U_{DS}), and specific choices for the claim differentiator J and its considered values j. Table 1 shows the mapping of our framework to standard group fairness criteria. For example, the acceptance rate is equivalent to the expected DS utility ($E(U_{DS})$) without any claim differentiator ($J = \emptyset$) if the utility weights are chosen as $w_{11} = w_{10} = 1$ and $w_{01} = w_{00} = 0$. Similarly, for $J = Y$, $j \in \{1\}$, and $w_{11} = w_{01} = 1$, the expected DS utility ($E(U_{DS}|Y = 1)$) corresponds to the true positive rate.

Table 2. Mapping of standard group fairness metrics to our utility-based approach under Egalitarianism

General conditions U_{DS} weights (for groups $a \in \{0,1\}$)	J	j	**Equivalent fairness criterion**
$w_{11} = w_{10} \neq w_{01} = w_{00} \ \wedge \ w_{dy} \perp a$	$\emptyset$	-	Statistical parity
$w_{11} = w_{10} \neq w_{01} = w_{00} \ \wedge \ w_{dy} \perp a$	L	l	Conditional statistical parity
$w_{11} \neq w_{01} \ \wedge \ w_{d1} \perp a$	Y	$\{1\}$	Equality of opportunity
$w_{10} \neq w_{00} \ \wedge \ w_{d0} \perp a$	Y	$\{0\}$	False positive rate parity
$w_{11} \neq w_{01} \wedge w_{10} \neq w_{00} \ \wedge \ w_{dy} \perp a$	Y	$\{0,1\}$	Equalized odds
$w_{11} \neq w_{10} \ \wedge \ w_{1y} \perp a$	D	$\{1\}$	Predictive parity
$w_{01} \neq w_{00} \ \wedge \ w_{0y} \perp a$	D	$\{0\}$	False omission rate parity
$w_{11} \neq w_{10} \wedge w_{01} \neq w_{00} \ \wedge \ w_{dy} \perp a$	D	$\{0,1\}$	Sufficiency

4.1 Standard group fairness criteria through the lens of our utility-based approach

The examples in Table 1 are not the only possibilities of utility matrices that lead to equivalence with a standard group fairness metric. Equality of U_{DS} between two relevant groups is insensitive against some changes of the utility matrix W. In particular, adding a constant to all matrix elements, or multiplying them with a constant factor, does not change the fairness criterion.[8] Thus, different utility matrices may lead to an equivalence to one of the standard group fairness criteria. In this section, we show under which conditions we achieve such an equivalence (see Table 2 for a summary of the results w.r.t.: (conditional) statistical parity,[9] equality of opportunity, false positive rate (FPR) parity, equalized odds, predictive parity, false omission rate (FOR) parity, and sufficiency.[10] The mathematical definitions of these criteria can be found in Table 3 in Appendix A. In the following, we focus on statistical parity, equality of opportunity, and predictive parity as prototypical examples. We refer the interested reader to the Appendix B.3 for a similar mapping of other standard group fairness criteria.

Statistical parity (also called demographic parity or group fairness [19]) is defined as $P(D = 1|A = 0) = P(D = 1|A = 1)$.

Proposition 4.1 (Statistical parity as utility-based fairness). *If the utility weights of all possible outcomes (as defined in Section 3.1) do not depend on the group membership ($w_{dy} \perp a$), and $w_{11} = w_{10} \neq w_{01} = w_{00}$, then the egalitarian pattern fairness condition with $J = \emptyset$ is equivalent to statistical parity.*

The formal proof of Proposition 4.1 can be found in Appendix B.1.1.

To measure the different degrees to which egalitarian fairness is fulfilled, we can introduce a quantitative fairness *metric* F. One option is to compute the absolute difference between the two groups' expected utilities:[11]

$$F_{\text{egalitarianism}} = |E(U_{DS}|J = j, A = 0) - E(U_{DS}|J = j, A = 1)| \tag{6}$$

Up to a multiplicative constant, this measure is equivalent to the degree to which statistical parity is fulfilled:

Corollary 4.2 (Partial fulfillment of statistical parity in terms of utility-based fairness). *Suppose that the degree to which statistical parity is fulfilled is defined as the absolute difference in decision ratios across groups, i.e.,*

[8] This allows for choosing a convenient reference point for utility, e.g. setting one of the elements to 0. By defining another one to have the value of 1, a scaling is introduced. See also [20] for a discussion of this topic.
[9] Notice that the claim differentiator for conditional statistical parity is defined as $J = L$, where L denotes legitimate attributes that can take the values l.
[10] Note that we focus on fairness criteria that are based on the decisions D and actual outcomes Y. However, this idea generalizes to fairness definitions that are based on predicted scores and actual outcomes, such as balance for the positive/negative class and well-calibration (see [13, 41, 65] for a definition of these criteria).
[11] Note that other metrics could be used, e.g., the ratio of the two expected utilities.

$|P(D = 1|A = 0) - P(D = 1|A = 1)|$. *If the utility weights do not depend on the group membership ($w_{dy} \perp a$), and $w_{11} = w_{10} \neq w_{01} = w_{00}$ (i.e., $w_{1y} \neq w_{0y}$), and $J = \emptyset$, then the degree to which egalitarianism is fulfilled is equivalent to the degree to which statistical parity is fulfilled, multiplied by $|w_{1y} - w_{0y}|$.*

The formal proof of Corollary 4.2 can be found in Appendix B.1.2.

Equality of opportunity (also called TPR parity) is defined as $P(D = 1|Y = 1, A = 0) = P(D = 1|Y = 1, A = 1)$, i.e., it requires parity of true positive rates (TPR) across groups $a \in A$ [23]. In this case, not all values of the claim differentiator J are considered to be relevant: we are only concerned with individuals of type $Y = 1$.

Proposition 4.3 (Equality of opportunity as utility-based fairness). *If w_{11} and w_{01} do not depend on the group membership ($w_{d1} \perp a$), and $w_{11} \neq w_{01}$, then the egalitarian pattern fairness condition with $J = Y$ and $j \in \{1\}$ is equivalent to equality of opportunity.*

The formal proof of Proposition 4.3 can be found in Appendix B.1.3. Compared to statistical parity, equality of opportunity only requires equal acceptance rates across those subgroups of A who are of type $Y = 1$. This corresponds to the claim differentiator $J = Y$ with $j \in \{1\}$. [12]

Predictive parity (also called PPV parity [7] or outcome test [63]) is defined as $P(Y = 1|D = 1, A = 0) = P(Y = 1|D = 1, A = 1)$. It requires parity of positive predictive value (PPV) rates across groups $a \in A$.

Proposition 4.4 (Predictive parity as utility-based fairness). *If w_{11} and w_{10} do not depend on the group membership ($w_{1y} \perp a$), and $w_{11} \neq w_{10}$, then the egalitarian pattern fairness condition with $J = D$ and $j \in \{1\}$ is equivalent to predictive parity.*

The formal proof of Proposition 4.4 can be found in Appendix B.1.4. [13, 14]

4.2 Uncovering the moral assumptions of standard group fairness metrics

Considering Table 2, we see that each standard group fairness criterion (a) constitutes a specific way of measuring the benefit/harm of decision subjects, (b) embeds assumptions about who has equal or different moral claims to utility, and (c) requires equality. All these elements correspond to normative choices that define what kind of fairness is achieved.

If we were to, for example, demand equality of opportunity for men and women in credit lending (where D is the bank's decision to either approve a loan ($D = 1$) or reject it ($D = 0$), and Y is the loan applicant's ability to repay the loan ($Y = 1$) or not ($Y = 0$)), we make the following assumptions: The benefit derived by being granted a loan is the same for each individual and the same for men and women. Only people who repay their loans have a legitimate claim to utility, and we don't need to consider the consequences for people who do not repay. Fairness means equalizing the acceptance rates of men and women of the morally relevant group (those who would repay a granted loan), even if this leads to undesirable outcomes for both men and women – other solutions are not considered.

[12] See Corollary B.1 in Appendix B.2 for the extension to a partial fulfillment of equality of opportunity.
[13] See Corollary B.2 in Appendix B.2 for the extension to a partial fulfillment of predictive parity.
[14] Notice that the fairness notion *well-calibration* is related to PPV parity but it is defined for scores instead of binary decisions: $P(Y = 1|S = s, A = 0) = P(Y = 1|S = s, A = 1)$. This requires that for each predicted score $s \in S$, individuals of all groups a have equal chances of belonging to the positive class [13]. Our proposed approach is equivalent to satisfying well-calibration if $J = S$, $w_{s1} = 1$, $w_{s0} = 0$, and using an egalitarian pattern of justice. In this case, the claim differentiator is the predicted score s, and all possible values of s need to be considered. Notice that, in this special case, the DS utility weights (denoted by w_{sy}) only depend on Y and are uniform across the entire range of scores. If one wants to extend well-calibration to take score-specific consequences of outcomes into account, this can be done easily by introducing score-specific utilities w_{sy}. The stronger definition of well-calibration ($P(Y = 1|S = s, A = 0) = P(Y = 1|S = s, A = 1) = s$), which is sometimes also called *calibration by groups* [5] or *calibration within groups* [41, 65], is equivalent to requiring that $\forall a \in A, \forall s \in S \;\; E(U_{DS}|J = s, A = a) = s$, for $J = S$, $w_{s1} = 1$, $w_{s0} = 0$. Here, the pattern is stronger than just egalitarianism.

All of these assumptions can be disputed for good reasons. For example, should we really ignore that being granted a loan might not only be beneficial for someone who cannot repay it? And is it morally acceptable to ignore the consequences for the defaulters? Also, is it really desirable to make every group worse off just for the sake of equality? These questions come up naturally when we analyze the utility matrix, the relevant groups, the claim differentiator, and the pattern of justice. Our framework shows possible alternatives for each component. This helps considerably to decide whether or not the chosen fairness criterion is morally appropriate and forces stakeholders to make their moral assumptions explicit, which are usually left implicit in standard approaches for choosing between group fairness criteria.

5 A SIMPLE APPLICATION EXAMPLE

Suppose that an ML-based decision-making system is used to identify those patients in a cancer population that will benefit from an innovative drug. Patients from the positive class ($Y = 1$) do not develop side effects after the drug treatment (or the side effects are negligible), i.e., they would benefit from the treatment because it cures their cancer. But those from the negative class ($Y = 0$) suffer from side effects of the cancer treatment. For the sake of this argument, let us assume that, despite being cured of cancer, those side effects require another treatment, which reduces life quality significantly over the next year. Due to the high cost of both treatments (the one against cancer and the one to treat the potential side effects), only individuals with a high likelihood of not developing any side effects (p) are treated ($D = 1$). More specifically, we assume that the optimal decision from the perspective of the decision makers (e.g., the hospital) would be to treat all individuals with a probability p that lies above 50% ($p = P(Y = 1) > 0.5$). We further assume that due to the non-representative selection of the research subjects for clinical trials, individuals from the minority group are much more likely to suffer from side effects (i.e., have lower probabilities p of not developing side effects). Absent any fairness considerations, this results in a lower treatment rate for the minority group. One might argue that the selection of cancer patients for treatment with the new drug should be made in a fair manner to avoid disadvantaging individuals in the minority group. This requires the elicitation of a morally appropriate group fairness metric. First, we will use established methods to select a standard metric. Then, we will apply our proposed utility-based approach. We will compare the results of both methods and analyze their implications.

First, using existing approaches to select one of the standard group fairness criteria [8, 27, 45], one might argue that statistical parity is an appropriate choice because the likelihood of requiring additional expensive treatment (due to developing side effects) does not determine how deserving people are to live without cancer – even if this may be a relevant consideration for efficiency reasons absent any fairness constraints. Thus, the chances of treatment should be equal for individuals of both groups.

Second, we elicit a morally appropriate fairness criterion by going through the four components of our framework: Regarding the definition of the relevant groups to compare, the example's assumption is that it is the minority group for which fairness should be ensured in comparison to the majority group. For example, it may be argued that this is reasonable due to the causal link between a patient's group membership and the likelihood of developing side effects (see Section 3.2). As for the claim differentiator, it may be assumed that all individuals have the same moral claim to utility, i.e., that there is no justifiable argument to differentiate between individuals' deservingness (or necessity or urgency) to be treated. In this case, following the same moral standpoint as above, there would not be any claim differentiator ($J = \emptyset$), equivalent to the case of statistical parity. However, critical differences may emerge when going through the other two steps of our framework, i.e., the evaluation of the utility of the DS (as introduced in Section 3.1) and the specification of the pattern for a just distribution of the utility derived by the cancer patients (see Section 3.4).

Let us now specify the DS utility. Using the disability-adjusted life years (DALY)[15] as a measure for patients' negative expected utilities to compare different outcomes of the medical treatment, we may specify the DS utility as follows: Individuals without any side effects receiving the treatment can live a cancer-free life, defining our reference point: $w_{11} = 0$ (representing zero DALYs). Individuals that do not receive the treatment continue living for one more year with the disease burden: $w_{00} = w_{01} = -0.4$ (representing slightly less than half a DALY). The utility of individuals developing side effects after having received the treatment depends on the severity of those side effects. For this simple example, we assume that the side effects are considerable but do not result in death. More precisely, we assume that the burden of the side effects and the additional treatment is equivalent to -0.8 DALYs (i.e., for the assumed year of life, $w_{10} = -0.8$). This is represented by the utility matrix in Fig. 1b, next to the DS utility matrix equivalent to the standard group fairness criteria statistical parity in Fig. 1a.

Next, we need to specify a pattern of justice, which defines what a just distribution looks like. We assume that maximizing the expected utility of the worst-off group (i.e., a maximin DS utility distribution) is desirable from a fairness perspective, as one might reasonably argue in a risk-averse health context [22], which would be in line with Rawls' initial original position [54]. This example shows that our general framework results in a different fairness metric. Not only is the benefit measured differently because we are taking the consequences of a decision (including possible side effects) into account, but we also apply a different pattern of justice.

We will now show that enforcing statistical parity does not necessarily make the minority group better off, on average. To ensure statistical parity, more individuals from the minority group have to be treated, compared to the unconstrained optimum, since minority individuals have a systematically lower p. However, whether being treated (i.e., switching from $D = 0$ to $D = 1$) is desirable for the patients, depends on the side effects: those who do not develop side effects gain utility (i.e., their expected utility changes from -0.4 to 0) and those who do develop side effects lose utility (i.e., their expected utility changes from -0.4 to -0.8). Apart from degenerate cases ($p = 0$ and $p = 1$), patients do not know with certainty if they will develop side effects, as the outcome Y is unknown. In expectation, a treatment is only desirable for individuals with $p >= \frac{w_{00} - w_{10}}{w_{11} - w_{01} - w_{10} + w_{00}} = 0.5$. Hence, increasing the number of treated minority patients is problematic, as the patients of the minority group who are treated additionally experience a disadvantage by the treatment rather than an advantage[16]. This is completely disregarded by the fairness metric statistical parity, which implicitly assumes that a positive decision is desirable for anyone ($w_{11} = w_{10} = 1$ and $w_{01} = w_{00} = 0$). In fact, in this scenario, enforcing statistical parity would likely make both groups worse off (by increasing/decreasing the number of treated patients in the minority/majority group), compared to the unconstrained case, in order to equalize the share of treated patients in the two groups – leading to a classical case of the "leveling down objection".

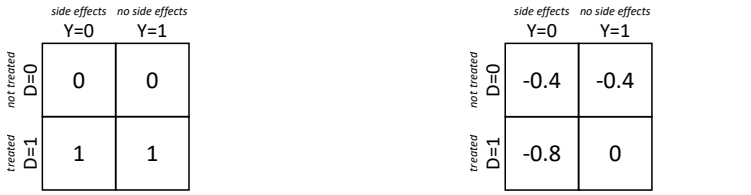
Applying the maximin pattern of justice, in contrast, can prevent us from producing 'fairness' at the cost of the minority group, which would contradict the overall goal of improving the situation for the minority group.

6 DISCUSSION

In this section, we discuss how our proposed framework alleviates the previously discussed limitations of standard group fairness criteria, and we comment on the limitations of our expanded definition of group fairness.

[15] DALY is a generic measure of disease burden calculated as the sum of the years of life lost (YLL) due to dying early and the years lost due to disability or disease (YLD), i.e., DALY = YLL + YLD, where one DALY represents the loss of the equivalent of one year of full health.

[16] Recall that, without fairness consideration, the hospital decided to treat patients with $p > 0.5$, due to cost considerations. Thus, increasing the number of treated patients in a group requires treating patients with $p < 0.5$



(a) Statistical parity (b) DS utility with considerable side effects

Fig. 1. Decision subject utilities for the medical example. Left, the DS utility matrix for the standard group fairness criteria statistical parity is shown. The matrix on the right represents the DS utility (in DALYs, here represented with negative utility) for the medical drug treatment example with considerable, burdensome side effects for the treated patients.

6.1 Alleviating limitations of existing fairness criteria

Standard group fairness criteria are special cases of our generalized group fairness framework. The suggested extension allows the alleviation of several of the standard group fairness limitations that we discussed in Section 2.1.

The "leveling down objection". The "leveling down objection" is a prevalent anti-egalitarianism argument [17, 53] saying that less inequality is not desirable if this requires lowering the better-off group's utility to match the one of the worse-off group. On this basis, choosing egalitarianism as the pattern of justice has been criticized in the algorithmic fairness literature (see, e.g., [32, 45, 66]). Our approach allows using other patterns of justice, such as maximin, prioritarianism, or sufficientarianism (see Section 3.4). Other patterns that can be formalized as mathematical constraints may also be used. One could, for example, combine several patterns into one and require equal expected utilities across groups as long as none of the groups is better off than it would be without any fairness requirement. This would represent a combination of egalitarianism and a group-specific baseline threshold (similar to sufficientarianism), making a "leveling down" of the better-off group impossible and adhering to the Pareto principle.

Focus on decisions instead of consequences. Standard group fairness criteria only consider the distribution of *either* D *or* Y. This can be interpreted as analyzing the distribution of utility but assuming that utility is equivalent to *either* D *or* Y instead of, for example, the combination of D *and* Y. Standard group fairness criteria thus represent a very confining definition of utility. Our approach acknowledges that the utility of the decision subjects may depend on a combination of different attributes such as one's ability to repay a loan or one's socioeconomic status (see, e.g., [10, 28, 66]. This is represented through the utility function described in Section 3.1, which can easily be extended (e.g., to take group-specific utility functions into account).

Limited set of fairness definitions. Previous attempts to guide stakeholders in choosing appropriate fairness criteria have taken on the form of explicit rules, such as in [46, 56, 57]. Works like [8, 27, 45] have provided unifying moral frameworks for understanding existing notions of algorithmic fairness, but they still presuppose a limited set of fairness definitions from which stakeholders can choose. While [27] consider the distribution of *undeserved* utility (what they call the *difference between an individual's actual and effort-based utility*), [45] and [8] use the decision subject utility U_{DS} to derive a morally appropriate group fairness definition. This is similar to the approach presented in this paper; however, they only consider two options $U_{DS} = D$ and $U_{DS} = Y$, while our approach allows for arbitrary functions f for the utility: $U_{DS} = f(D, Y)$. Furthermore, [8, 27, 45] only consider egalitarian notions of fairness, and it remains unclear how non-egalitarian notions of fairness fit in.

As discussed in Section 2.2, previous works have expanded standard group fairness metrics. However, the resulting fairness notions diverge from standard metrics on some of the four components in our framework. The other components are held constant compared to standard group fairness metrics, while the assumptions that are encompassed in the choice to keep these components constant are not made explicit. Therefore, the criteria resulting from these expansions are still somewhat limited. We provide a method that integrates these prior works in a unifying framework and link the different choices to morally relevant concepts with respect to the utility function for decision subjects (Section 3.1), the relevant groups to compare (Section 3.2), the subgroups with equal claims to utility (Section 3.3), and the pattern for a just distribution of utility (Section 3.4).

6.2 Limitations

Fundamental assumptions of standard group fairness criteria. While our framework extends standard group fairness criteria, we still share some of the fundamental assumptions embedded in group fairness. First, we compare averages across groups, which has been criticized for being vulnerable to fairness gerrymandering [19]. There could be systematic differences between groups despite them having the same averages, e.g., due to a different distribution within groups. Second, one could criticize that we (and group fairness notions in general) cannot distinguish cases in which membership in the relevant groups has a causal influence on the outcomes and decisions or whether they just happen to be correlated – in cases where both result in the same distribution of utilities. Contrary to that, counterfactual fairness [43] demands that the sensitive attribute (and its proxies) do not influence the final decision. While we cannot guarantee that our group fairness criteria would fulfill such a strict requirement, we argue that our approach to group fairness most likely avoids the objection of fairness gerrymandering and causal irrelevance in practice. Our practical solution to these objections is to require at least a weak causal link for the specification of relevant groups (as mentioned in Section 3.2): We demand that individuals belong to a relevant group that is likely to be the cause of an unjust inequality. This way, we reduce the probability that group fairness is evaluated in a situation where the inequality is caused by spurious unfortunate correlations. When defining the relevant groups, we could make them increasingly narrow. This idea of increasingly narrow groups aligns with the concept of multicalibration, which is motivated by the concept of individual fairness. It can be seen as a further extension of well-calibration. Multicalibration calls for calibrating every efficiently-identifiable subgroup R of a computationally-identifiable subset C of the population P [25]. Intuitively, multicalibration can also be seen as a special case of our proposed framework, where $\forall r \in R \;\; E(U_{DS}|J = s, R = r) = s$ for all subgroups $R \in C$, for $J = S$, $w_{s1} = 1$, $w_{s0} = 0$. However, the larger the number of subgroups, the more difficult it becomes to make moral judgments about them. Therefore, instead of only considering subgroups based on computational efficiency as in multicalibration, we focus on groups that meet the weak causality requirement. Furthermore, w.r.t. multicalibration, our proposed framework demonstrates that benefits/harms are measured in a narrow way ($w_{s1} = 1$, $w_{s0} = 0$), which can be extended using the flexibility of the DS utility function. Our framework thus again reveals the normative choices of these fairness notions.

Economic notions of fairness. As we explained in Section 2.2, our framework builds on existing extensions of standard group fairness metrics and tries to structure these. Yet, there are still some extensions that do not fit neatly into our framework. As far as we are aware, this mainly concerns Zafar et al. [68]'s interpretation of group-level envy-freeness for fair machine learning. Contrary to Hossain et al. [31], they postulate that group-level envy-freeness is fulfilled if "every sensitive attribute group (e.g., men and women) prefers the set of decisions they receive over the set of decisions they would have received had they collectively presented themselves to the system as members of a different sensitive group"

[68, p. 3]. This fairness criterion is structurally different from the fairness criteria in our framework: Our fairness criteria compare the average utilities of different groups. Instead, the envy-freeness criterion compares the average utility of a single group to the expected utility of this group if this group had a different sensitive attribute – it thus compares the average utility of a single group under different assumptions. Similarly, Kim et al. [38]'s preference-informed statistical parity compares utilities of groups across alternative classifiers instead of comparing utilities between groups for a single classifier. The question is thus not about how a classifier should distribute utilities between equally deserving groups but about whether a classifier makes every group better off than some alternative.

Theories of distributive justice. While our approach creates a link between group fairness and different theories of justice, it does not cover theories of distributive justice that are structurally different from the ones we discussed, e.g., Nozick's entitlement theory [51]. It is unclear how such theories could be represented in formalized fairness criteria.

Utility in practice. While we showcased a simplified approach for specifying utility matrices in Section 5, we recognize that defining a utility function is difficult in practice [20, 61]. Moreover, we only presented a utility function that is linear in Y and D. Our framework allows for more complex utility functions, but these are even harder to define. We describe how utility functions can be defined through the lens of a simplified medical example. However, determining how to quantify the utility of decisions in general (i.e., using a clearly defined guideline that is applicable in any application context, which might require an empirical approach), falls outside the scope of this paper. Another limitation is that we only proposed simple metrics derived from the utility matrix but no combination of these (e.g., separation as the combination of parity in true positive rates and false positive rates). While we could represent these combined metrics in our framework, it is again not obvious what the best way to do so is. Here, we refer to [5] to see how information theory's concept of mutual information can be used to represent separation and sufficiency.

7 CONCLUSION

In this paper, we have proposed a novel generalized definition of group fairness that is based on a comprehensive framework that unifies and extends existing work on what can broadly be described as "group fairness". As part of this, we have also suggested a new definition of group fairness as a category of metrics that are concerned with the just distribution of utility among relevant groups. Our framework consists of four components: (1) utility of the decision subjects, (2) relevant groups to compare, (3) claim differentiator to derive subgroups to compare that matter, and (4) patterns for a just distribution of utility. These components form a lens through which we can interpret existing fairness metrics. The main benefits of our framework are that it allows us to decode the normative choices hidden in fairness criteria and that it yields a structured way of creating unique and context-sensitive fairness criteria. Using a simple example, we showed that for different versions of prediction-based decision making systems, our approach can determine the fairest solution, according to the chosen normative choices. However, the question of how a fair solution can be achieved optimally remains open. More research is needed to incorporate our novel understanding of group fairness into automated decision making systems, for example, using pre-processing [34, 35], in-processing [36, 48, 69], or post-processing techniques [7, 15, 23].

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. 2023. A Justice-Based Framework for the Analysis of Algorithmic Fairness-Utility Trade-Offs. (2023). Unpublished manuscript – submitted as part of the supplementary material.

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

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

[6] Solon Barocas and Andrew D Selbst. 2016. Big Data's Disparate Impact. *California Law Review* 104, 3 (2016), 671–732. http://www.jstor.org/stable/24758720

[7] Joachim Baumann, Anikó Hannák, 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 (FAccT '22)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3531146.3534645

[8] 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 (SDS)*. 19–25. https://doi.org/10.1109/SDS54800.2022.00011

[9] Omer Ben-Porat, Fedor Sandomirskiy, and Moshe Tennenholtz. 2021. Protecting the protected group: Circumventing harmful fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 5176–5184.

[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] Reuben Binns. 2020. On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 514–524.

[12] 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.

[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] 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

[15] 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.

[16] Kate Crawford. 2016. Artificial intelligence's white guy problem. *The New York Times* 25, 06 (2016).

[17] Roger Crisp. 2003. Equality, Priority, and Compassion. 113, 4 (2003), 745–763. https://doi.org/10.1086/373954

[18] Emily Diana, Wesley Gill, Michael Kearns, Krishnaram Kenthapadi, and Aaron Roth. 2021. Minimax group fairness: Algorithms and experiments. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 66–76.

[19] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

[20] Charles Elkan. 2001. The Foundations of Cost-Sensitive Learning. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2 (IJCAI'01)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 973–978.

[21] Jessie Finocchiaro, Roland Maio, Faidra Monachou, Gourab K Patro, Manish Raghavan, Ana-Andreea Stoica, and Stratis Tsirtsis. 2021. Bridging machine learning and mechanism design towards algorithmic fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 489–503.

[22] Ulrik Franke. 2021. Rawls's Original Position and Algorithmic Fairness. *Philosophy & Technology* 34, 4 (2021), 1803–1817. https://doi.org/10.1007/s13347-021-00488-x

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

[24] 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/

[25] Ursula Hebert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. 2018. Multicalibration: Calibration for the (Computationally-Identifiable) Masses. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer Dy and Andreas Krause (Eds.). PMLR, 1939–1948. https://proceedings.mlr.press/v80/hebert-johnson18a.html

[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] Sune Holm. 2022. The Fairness in Algorithmic Fairness. *Res Publica* (2022), 1–17.

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

[31] Safwan Hossain, Andjela Mladenovic, and Nisarg Shah. 2020. Designing fairly fair classifiers via economic fairness notions. In *Proceedings of The Web Conference 2020*. 1559–1569.

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

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

[34] Faisal Kamiran and Toon Calders. 2009. Classifying without discriminating. In *2009 2nd International Conference on Computer, Control and Communication*. 1–6. https://doi.org/10.1109/IC4.2009.4909197

[35] Faisal Kamiran and Toon Calders. 2012. Data Preprocessing Techniques for Classification without Discrimination. *Knowl. Inf. Syst.* 33, 1 (oct 2012), 1–33. https://doi.org/10.1007/s10115-011-0463-8

[36] Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. 2011. Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*. IEEE, 643–650.

[37] Maximilian Kasy and Rediet Abebe. 2021. Fairness, equality, and power in algorithmic decision-making. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 576–586.

[38] Michael P. Kim, Aleksandra Korolova, Guy N. Rothblum, and Gal Yona. 2019. Preference-Informed Fairness. *CoRR* abs/1904.01793 (2019). arXiv:1904.01793 http://arxiv.org/abs/1904.01793

[39] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. *Human Decisions and Machine Predictions*. Working Paper 23180. National Bureau of Economic Research. https://doi.org/10.3386/w23180

[40] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Cass R Sunstein. 2019. Discrimination in the Age of Algorithms. *Journal of Legal Analysis* 10 (2019), 113–174. https://doi.org/10.1093/jla/laz001

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

[42] Matthias Kuppler, Christoph Kern, Ruben L. Bach, and Frauke Kreuter. 2021. Distributive Justice and Fairness Metrics in Automated Decision-making: How Much Overlap Is There? arXiv:2105.01441 [stat.ML]

[43] Matt J Kusner, Joshua R Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. *arXiv preprint arXiv:1703.06856* (2017).

[44] Christian List. 2022. Social Choice Theory. In *The Stanford Encyclopedia of Philosophy* (Spring 2022 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.

[45] 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.

[46] 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

[47] Natalia Martinez, Martin Bertran, and Guillermo Sapiro. 2020. Minimax pareto fairness: A multi objective perspective. In *International Conference on Machine Learning*. PMLR, 6755–6764.

[48] 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.

[49] 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.

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

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

[52] 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.

[53] Derek Parfit. 1995. *Equality or priority*. Department of Philosophy, University of Kansas.

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

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

[56] Boris Ruf and Marcin Detyniecki. 2022. A Tool Bundle for AI Fairness in Practice. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–3.

[57] 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).

[58] Aaron Sankin, Dhruv Mehrotra, Surya Mattu, and Annie Gilbertson. 2021. Crime Prediction Software Promised to Be Free of Biases. New Data Shows It Perpetuates Them. *The Markup* (2021). https://themarkup.org/prediction-bias/2021/12/02/crime-prediction-software-promised-to-be-free-of-biases-new-data-shows-it-perpetuates-them

[59] 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.

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

[61] 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

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

[63] Camelia Simoiu, Sam Corbett-Davies, Sharad Goel, et al. 2017. The problem of infra-marginality in outcome tests for discrimination. *The Annals of Applied Statistics* 11, 3 (2017), 1193–1216.

[64] Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P. Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. 2018. A Unified Approach to Quantifying Algorithmic Unfairness: Measuring Individual & Group Unfairness via Inequality Indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (London, United Kingdom) *(KDD '18)*. Association for Computing Machinery, New York, NY, USA, 2239–2248. https://doi.org/10.1145/3219819.3220046

[65] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)*. IEEE, 1–7.

[66] 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).

[67] Pak-Hang Wong. 2020. Democratizing Algorithmic Fairness. *Philosophy & Technology* 33, 2 (2020), 225–244. https://doi.org/10.1007/s13347-019-00355-w

[68] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, Krishna P. Gummadi, and Adrian Weller. 2017. From Parity to Preference-Based Notions of Fairness in Classification. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 228–238.

[69] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. 2017. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*. PMLR, 962–970.

A STANDARD GROUP FAIRNESS CRITERIA

Here, we briefly introduce the most discussed group fairness criteria. Table 3 list the parity requirements associated with these criteria. *Statistical parity* demands that the share of positive decisions is equal between socio-demographic groups (defined by the sensitive attribute $A = \{0, 1\}$) [19] – this is only required for a set of so-called legitimate attributes $l \in L$ for the criterion *conditional statistical parity* [15]. *Equality of opportunity*, similarly, demands equal shares of positive decisions between socio-demographic groups, but only for those whose target variable is positive ($Y = 1$) [23] – thus, it is sometimes also referred to as true positive rate (TPR) parity. *Equalized odds* – sometimes also called *separation* – requires both equality of opportunity and FPR parity (which is similar to equality of opportunity, however, it is limited to individuals of type $Y = 0$). In contrast, *predictive parity* demands equal shares of individuals of type $Y = 1$ across socio-demographic groups, but only for those who received a positive decision $D = 1$ [7] – thus, it is sometimes also referred to as positive predictive value (PPV) parity. *Sufficiency* requires both PPV parity and false omission rate (FOR) parity (which is similar to PPV parity, however, it is limited to individuals who received a negative decision $D = 0$).

B MAPPING STANDARD GROUP FAIRNESS CRITERIA TO OUR UTILITY-BASED APPROACH

B.1 Omitted proofs

B.1.1 Proof of Proposition 4.1. Recall that the utility-based fairness following the pattern of egalitarianism requires equal expected utilities between groups:

$$E(U_{DS}|J = j, A = 0) = E(U_{DS}|J = j, A = 1) \tag{B.7}$$

Table 3. Standard group fairness criteria

Fairness criterion	**Parity requirement**
Statistical parity	$P(D=1\|A=0) = P(D=1\|A=1)$
Conditional statistical parity	$P(D=1\|L=l, A=0) = P(D=1\|L=l, A=1)$
Equality of opportunity	$P(D=1\|Y=1, A=0) = P(D=1\|Y=1, A=1)$
False positive rate parity	$P(D=1\|Y=0, A=0) = P(D=1\|Y=0, A=1)$
Equalized odds	$P(D=1\|Y=y, A=0) = P(D=1\|Y=y, A=1)$, for $y \in \{0,1\}$
Predictive parity	$P(Y=1\|D=1, A=0) = P(Y=1\|D=1, A=1)$
False omission rate parity	$P(Y=1\|D=0, A=0) = P(Y=1\|D=0, A=1)$
Sufficiency	$P(Y=1\|D=d, A=0) = P(Y=1\|D=d, A=1)$, for $d \in \{0,1\}$

Since there is no claim differentiator (i.e., $J = \emptyset$), this can be simplified to:

$$E(U_{DS}|A=0) = E(U_{DS}|A=1) \tag{B.8}$$

For $w_{11} = w_{10}$ and $w_{01} = w_{00}$, the decision subject utility (see Equation 1) is:

$$u_{DS,i} = w_{0y} + (w_{1y} - w_{0y}) \cdot d_i, \tag{B.9}$$

where w_{1y} denotes the decision subject utility associated with a positive decision ($D=1$) and w_{0y} denotes the decision subject utility associated with a negative decision ($D=0$). Thus, the expected utility for individuals of group a can be written as:

$$E(U_{DS}|A=a) = w_{0y} + (w_{1y} - w_{0y}) \cdot P(D=1|A=a). \tag{B.10}$$

If the utility weights of all possible outcomes do not depend on the group membership ($w_{dy} \perp a$), and $w_{1y} \neq w_{0y}$[17], then the utility-based fairness following the pattern of egalitarianism (see Equation B.8) requires:

$$\begin{aligned} w_{0y} + (w_{1y} - w_{0y}) \cdot P(D=1|A=0) &= w_{0y} + (w_{1y} - w_{0y}) \cdot P(D=1|A=1) \\ \Leftrightarrow (w_{1y} - w_{0y}) \cdot P(D=1|A=0) &= (w_{1y} - w_{0y}) \cdot P(D=1|A=1) \\ \Leftrightarrow P(D=1|A=0) &= P(D=1|A=1), \end{aligned} \tag{B.11}$$

where the last line is identical to statistical parity.

B.1.2 Proof of Corollary 4.2. Recall that the degree to which egalitarianism is fulfilled is defined as $F_{\text{egalitarianism}} = |E(U_{DS}|J=j, A=0) - E(U_{DS}|J=j, A=1)|$ (see Equation 6). If the utility weights of all possible outcomes do not depend on the group membership ($w_{dy} \perp a$), and $w_{11} = w_{10} \neq w_{01} = w_{00}$ (i.e., $w_{1y} \neq w_{0y}$), $J = \emptyset$, this can be written as (see Equations B.8 and B.10):

$$\begin{aligned} F_{\text{egalitarianism}} &= \big| \big(w_{0y} + (w_{1y} - w_{0y}) \cdot P(D=1|A=0)\big) \\ &\quad - \big(w_{0y} + (w_{1y} - w_{0y}) \cdot P(D=1|A=1)\big) \big| \\ &= \big| \big((w_{1y} - w_{0y}) \cdot P(D=1|A=0)\big) - \big((w_{1y} - w_{0y}) \cdot P(D=1|A=1)\big) \big| \\ &= |(w_{1y} - w_{0y}) \cdot (P(D=1|A=0) - P(D=1|A=1))| \end{aligned} \tag{B.12}$$

where the last line corresponds to a multiplication of $|w_{1y} - w_{0y}|$ with the degree to which statistical parity is fulfilled.

[17] If $w_{1y} = w_{0y}$, then the utility-based fairness following the pattern of egalitarianism would always be satisfied and the equivalence to statistical parity would not hold.

B.1.3 Proof of Proposition 4.3. Recall that the utility-based fairness following the pattern of egalitarianism requires equal expected utilities between groups:

$$E(U_{DS}|J = j, A = 0) = E(U_{DS}|J = j, A = 1) \tag{B.13}$$

Since the claim differentiator is the same as the attribute $Y = 1$, i.e., $J = Y$ and the only morally relevant value of Y is 1 (i.e., $j \in \{1\}$), this can be simplified to:

$$E(U_{DS}|Y = 1, A = 0) = E(U_{DS}|Y = 1, A = 1) \tag{B.14}$$

For $y_i = 1$, the decision subject utility (see Equation 1) is:

$$u_{DS,i} = w_{01} + (w_{11} - w_{01}) \cdot d_i. \tag{B.15}$$

Thus, the expected utility for individuals of type $Y = 1$ in group a can be written as:

$$E(U_{DS}|Y = 1, A = a) = w_{01} + (w_{11} - w_{01}) \cdot P(D = 1|Y = 1, A = a). \tag{B.16}$$

If w_{11} and w_{01} do not depend on the group membership ($w_{d1} \perp a$), and $w_{11} \neq w_{01}$[18], then the utility-based fairness following the pattern of egalitarianism (see Equation B.14) requires:

$$\begin{aligned} &w_{01} + (w_{11} - w_{01}) \cdot P(D = 1|Y = 1, A = 0) = w_{01} + (w_{11} - w_{01}) \cdot P(D = 1|Y = 1, A = 1) \\ \Leftrightarrow\ &(w_{11} - w_{01}) \cdot P(D = 1|Y = 1, A = 0) = (w_{11} - w_{01}) \cdot P(D = 1|Y = 1, A = 1) \\ \Leftrightarrow\ &P(D = 1|Y = 1, A = 0) = P(D = 1|Y = 1, A = 1), \end{aligned} \tag{B.17}$$

where the last line is identical to equality of opportunity.

B.1.4 Proof of Proposition 4.4. Recall that the utility-based fairness following the pattern of egalitarianism requires equal expected utilities between groups:

$$E(U_{DS}|J = j, A = 0) = E(U_{DS}|J = j, A = 1) \tag{B.18}$$

Since the claim differentiator is the same as the decision $D = 1$, i.e., $J = D$ and the only morally relevant value of D is 1 (i.e., $j \in \{1\}$), this can be simplified to:

$$E(U_{DS}|D = 1, A = 0) = E(U_{DS}|D = 1, A = 1) \tag{B.19}$$

For $d_i = 1$, the decision subject utility (see Equation 1) is:

$$u_{DS,i} = w_{10} + (w_{11} - w_{10}) \cdot y_i. \tag{B.20}$$

Thus, the expected utility for individuals in group a that are assigned the decision $D = 1$ can be written as:

$$E(U_{DS}|D = 1, A = a) = w_{10} + (w_{11} - w_{10}) \cdot P(Y = 1|D = 1, A = a). \tag{B.21}$$

[18] If $w_{11} = w_{01}$, then the utility-based fairness following the pattern of egalitarianism would always be satisfied and the equivalence to equality of opportunity would not hold.

If w_{11} and w_{10} do not depend on the group membership ($w_{1y} \perp a$), and $w_{11} \neq w_{10}$[19], then the utility-based fairness following the pattern of egalitarianism (see Equation B.19) requires:

$$\begin{aligned} & w_{10} + (w_{11} - w_{10}) \cdot P(Y = 1|D = 1, A = 0) = w_{10} + (w_{11} - w_{10}) \cdot P(Y = 1|D = 1, A = 1) \\ \Leftrightarrow\ & (w_{11} - w_{10}) \cdot P(Y = 1|D = 1, A = 0) = (w_{11} - w_{10}) \cdot P(Y = 1|D = 1, A = 1) \\ \Leftrightarrow\ & P(Y = 1|D = 1, A = 0) = P(Y = 1|D = 1, A = 1), \end{aligned} \tag{B.22}$$

where the last line is identical to predictive parity.

B.2 Additional corollaries

Let us consider the partial fulfillment of equality of opportunity, following Proposition 4.3. As is the case for statistical parity, there are differences when looking at the degree to which the two notions of fairness are fulfilled (equality of opportunity and the utility-based fairness under the conditions specified in Proposition 4.3)

Corollary B.1 (Partial fulfillment of equality of opportunity in terms of utility-based fairness). *Suppose that the degree to which equality of opportunity is fulfilled is defined as the absolute difference in decision ratios for individuals of type $Y = 1$ across groups, i.e., $|P(D = 1|Y = 1, A = 0) - P(D = 1|Y = 1, A = 1)|$. If w_{11} and w_{01} do not depend on the group membership ($w_{d1} \perp a$), $w_{11} \neq w_{01}$, $J = Y$, and $j \in \{1\}$, then the degree to which egalitarianism is fulfilled is equivalent to the degree to which equality of opportunity is fulfilled, multiplied by $|(w_{11} - w_{01})|$.*

Proof. Recall that the degree to which egalitarianism is fulfilled is defined as $F_{\text{egalitarianism}} = |E(U_{DS}|J = j, A = 0) - E(U_{DS}|J = j, A = 1)|$ (see Equation 6). If w_{11} and w_{01} do not depend on the group membership ($w_{d1} \perp a$), $w_{11} \neq w_{01}$, $J = Y$, and $j \in \{1\}$, this can be written as (see Equations B.14 and B.16):

$$\begin{aligned} F_{\text{egalitarianism}} &= |\ (w_{01} + (w_{11} - w_{01}) \cdot P(D = 1|Y = 1, A = 0)) \\ &\quad - (w_{01} + (w_{11} - w_{01}) \cdot P(D = 1|Y = 1, A = 1))\ | \\ &= |\ ((w_{11} - w_{01}) \cdot P(D = 1|Y = 1, A = 0)) \\ &\quad - ((w_{11} - w_{01}) \cdot P(D = 1|Y = 1, A = 1))\ | \\ &= |(w_{11} - w_{01}) \cdot (P(D = 1|Y = 1, A = 0) - P(D = 1|Y = 1, A = 1))\ | \end{aligned} \tag{B.23}$$

where the last line corresponds to a multiplication of $|w_{11} - w_{01}|$ with the degree to which equality of opportunity is fulfilled. □

As is the case for the other group fairness criteria, there are differences regarding the degree to which the two fairness notions of predictive parity and the utility-based fairness under the conditions specified in Proposition 4.4 are fulfilled:

Corollary B.2 (Partial fulfillment of predictive parity in terms of utility-based fairness). *Suppose that the degree to which predictive parity is fulfilled is defined as the absolute difference in the ratio of individuals that are of type $Y = 1$ among all those that are assigned the decision $D = 1$ across groups, i.e., $|P(Y = 1|D = 1, A = 0) - P(Y = 1|D = 1, A = 1)|$. If w_{11} and w_{10} do not depend on the group membership ($w_{1y} \perp a$), $w_{11} \neq w_{10}$, $J = D$, and $j \in \{1\}$, then the degree to which egalitarianism is fulfilled is equivalent to the degree to which predictive parity is fulfilled, multiplied by $|w_{11} - w_{10}|$.*

[19] If $w_{11} = w_{10}$, then the utility-based fairness following the pattern of egalitarianism would always be satisfied and the equivalence to predictive parity would not hold.

Proof. Recall that the degree to which egalitarianism is fulfilled is defined as $F_{\text{egalitarianism}} = |E(U_{DS}|J = j, A = 0) - E(U_{DS}|J = j, A = 1)|$ (see Equation 6). If w_{11} and w_{10} do not depend on the group membership ($w_{1y} \perp a$), $w_{11} \neq w_{10}$, $J = D$, and $j \in \{1\}$, this can be written as (see Equations B.19 and B.21):

$$\begin{aligned} F_{\text{egalitarianism}} &= | \, (w_{10} + (w_{11} - w_{10}) \cdot P(Y = 1|D = 1, A = 0)) \\ &\quad - (w_{10} + (w_{11} - w_{10}) \cdot P(Y = 1|D = 1, A = 1)) \, | \\ &= | \, ((w_{11} - w_{10}) \cdot P(Y = 1|D = 1, A = 0)) \\ &\quad - ((w_{11} - w_{10}) \cdot P(Y = 1|D = 1, A = 1)) \, | \\ &= |(w_{11} - w_{10}) \cdot (P(Y = 1|D = 1, A = 0) - P(Y = 1|D = 1, A = 1)) \, | \end{aligned} \tag{B.24}$$

where the last line corresponds to a multiplication of $|w_{11} - w_{10}|$ with the degree to which predictive parity is fulfilled. □

B.3 Mapping to other group fairness criteria

In Section 4, we mapped our utility-based approach to the three group fairness criteria statistical parity, equality of opportunity, and predictive parity. Here, we additionally show under which conditions our utility-based approach is equivalent to other group fairness criteria: conditional statistical parity, false positive rate parity, equalized odds, false omission rate parity, and sufficiency.

B.3.1 Conditional statistical parity. Conditional statistical parity is defined as $P(D = 1|L = l, A = 0) = P(D = 1|L = l, A = 1)$, where L is what [15] refer to as the *legitimate* attributes. Thus, conditional statistical parity requires equality of acceptance rates across all subgroups in $A = 0$ and $A = 1$ who are equal in their value l for L, where L can be any (combination of) feature(s) besides D and A.

Proposition B.3 (Conditional statistical parity as utility-based fairness). *If the utility weights of all possible outcomes do not depend on the group membership ($w_{dy} \perp a$), and $w_{11} = w_{10} \neq w_{01} = w_{00}$, then the egalitarian pattern fairness condition with $J = L$ is equivalent to conditional statistical parity.*

The proof of Proposition B.3 is similar to the one of Proposition 4.1.

Under these conditions, the degree to which $F_{\text{egalitarianism}}$ is fulfilled is equivalent to the degree to which conditional statistical parity is fulfilled, multiplied by $|w_{1y} - w_{0y}|$. This could easily be proved – similar to the proof of Corollary 4.2 but with the conditions of the utility-based fairness stated in Proposition B.3.

B.3.2 False positive rate (FPR) parity. FPR parity (also called predictive equality [15]) is defined as $P(D = 1|Y = 0, A = 0) = P(D = 1|Y = 0, A = 1)$, i.e., it requires parity of false positive rates (FPR) across groups $a \in A$.

Proposition B.4 (FPR parity as utility-based fairness). *If w_{10} and w_{00} do not depend on the group membership ($w_{d0} \perp a$), and $w_{10} \neq w_{00}$, then the egalitarian pattern fairness condition with $J = Y$ and $j \in \{0\}$ is equivalent to FPR parity.*

For $y_i = 0$, the decision subject utility (see Equation 1) is:

$$u_{DS,i} = w_{00} + (w_{10} - w_{00}) \cdot d_i. \tag{B.25}$$

Thus, the expected utility for individuals of type $Y = 0$ in group a can be written as:

$$E(U_{DS}|Y = 0, A = a) = w_{00} + (w_{10} - w_{00}) \cdot P(D = 1|Y = 0, A = a). \tag{B.26}$$

Hence, we simply require the utility weights w_{10} and w_{00} to be unequal and independent of a. Then, the proof of Proposition B.4 is similar to the one of Proposition 4.3.

If w_{10} and w_{00} do not depend on the group membership ($w_{d0} \perp a$), and $w_{10} \neq w_{00}$, then the degree to which $F_{\text{egalitarianism}}$ is fulfilled is equivalent to the degree to which FPR parity is fulfilled, multiplied by $|w_{10} - w_{00}|$. This could easily be proved – similar to the proof of Corollary B.1.

B.3.3 Equalized odds. Equalized odds (sometimes also referred to as separation [5]) is defined as $P(D = 1|Y = y, A = 0) = P(D = 1|Y = y, A = 1)$, for $y \in \{0, 1\}$.

Proposition B.5 (Equalized odds as utility-based fairness). *If the utility weights of all possible outcomes do not depend on the group membership ($w_{dy} \perp a$), $w_{11} \neq w_{01}$, and $w_{10} \neq w_{00}$, then the egalitarian pattern fairness condition with $J = Y$ and $j \in \{0, 1\}$ is equivalent to equalized odds.*

The conditions under which the utility-based fairness criteria is equivalent is shown separately for equality of opportunity (see Proposition 4.3) and FPR parity (see Proposition B.4). Since equalized odds requires equality of opportunity and FPR parity, the the conditions for both fairness criteria must be met (i.e., $w_{dy} \perp a$), $w_{11} \neq w_{01}$, $w_{10} \neq w_{00}$, $J = Y$, and $j \in \{0, 1\}$), so that the utility-based fairness constraint is equivalent to equalized odds.

B.3.4 False omission rate (FOR) parity. FOR parity is defined as $P(Y = 1|D = 0, A = 0) = P(Y = 1|D = 0, A = 1)$, i.e., it requires parity of false omission rates (FOR) across groups $a \in A$.

Proposition B.6 (FOR parity as utility-based fairness). *If w_{01} and w_{00} do not depend on the group membership ($w_{0y} \perp a$), and $w_{01} \neq w_{00}$, then the egalitarian pattern fairness condition with $J = D$, and $j \in \{0\}$ is equivalent to FOR parity.*

For $d_i = 0$, the decision subject utility (see Equation 1) is:

$$u_{DS,i} = w_{00} + (w_{01} - w_{00}) \cdot y_i. \tag{B.27}$$

Thus, the expected utility for individuals in group a that are assigned the decision $D = 0$ can be written as:

$$E(U_{DS}|D = 0, A = a) = w_{00} + (w_{01} - w_{00}) \cdot P(Y = 1|D = 0, A = a). \tag{B.28}$$

Hence, we simply require the utility weights w_{01} and w_{00} to be unequal and independent of a. Then, the proof of Proposition B.6 is similar to the one of Proposition 4.4.

If w_{01} and w_{00} do not depend on the group membership ($w_{0y} \perp a$), and $w_{01} \neq w_{00}$, then the degree to which $F_{\text{egalitarianism}}$ is fulfilled is equivalent to the degree to which FoR parity is fulfilled, multiplied by $|w_{01} - w_{00}|$. This could easily be proved – similar to the proof of Corollary B.2.

B.3.5 Sufficiency. Sufficiency is defined as $P(Y = 1|D = d, A = 0) = P(Y = 1|D = d, A = 1)$, for $d \in \{0, 1\}$ [5].

Proposition B.7 (Sufficiency as utility-based fairness). *If the utility weights of all possible outcomes do not depend on the group membership ($w_{dy} \perp a$), $w_{11} \neq w_{10}$, and $w_{01} \neq w_{00}$, then the egalitarian pattern fairness condition with $J = D$ and $j \in \{0, 1\}$ is equivalent to sufficiency.*

The conditions under which the utility-based fairness criteria is equivalent is shown separately for predictive parity (see Proposition 4.4) and FOR parity (see Proposition B.6). Since sufficiency requires predictive parity and FOR parity,

the the conditions for both fairness criteria must be met (i.e., $w_{dy} \perp a$), $w_{11} \neq w_{10}$, $w_{01} \neq w_{00}$, $J = D$, and $j \in \{0, 1\}$), so that the utility-based fairness constraint is equivalent to sufficiency.

C ANONYMOUS VERSION OF THE PAPER TITLED "A JUSTICE-BASED FRAMEWORK FOR THE ANALYSIS OF ALGORITHMIC FAIRNESS-UTILITY TRADE-OFFS"

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. However, these values are often hidden in the technicalities of the implementation of the decision-making system. In this paper, we propose a framework to make these value-laden choices clearly visible. We focus on a setting in which we want to find decision rules that balance the perspective of the decision maker and of 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 well-known theories of distributive justice and on the algorithmic literature to ask what a fair distribution of utility (or welfare) looks like. This allows us to derive a fairness score that we then compare to the decision maker's utility. As we focus on a setting in which we are given a trained model and have to choose a decision rule, we use the concept of Pareto efficiency to compare decision rules. Our proposed framework can both guide the implementation of a decision-making system and help with audits as it allows us to resurface the values implemented in a 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). 2023. A Justice-Based Framework for the Analysis of Algorithmic Fairness-Utility Trade-Offs. In *2023 ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '23), October 30 – November 1, 2023, Boston, US.* ACM, New York, NY, USA, 17 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

The increasing use of prediction-based decision-making systems has shown that this can easily lead to disadvantages for marginalized groups (see, e.g., [3, 13, 18, 21, 28, 52]). These systems are very unlikely to achieve fairness because they are optimized for goals *other than fairness*. Our framing hypothesis is that, beside pursing the decision maker's goal (e.g., to be as efficient or profitable as possible), a decision-making process should be fair towards the decision subjects, i.e., towards the individuals affected by the decisions. Often, these two goals conflict [39].[1] Navigating this trade-off requires making the values of the decision maker and the decision subjects explicit — to the point where they can be expressed as mathematical formulas. The perspective of fairness has been discussed in both computer science, coming up with many different so-called "fairness metrics" [49, 50] and, for a much longer time, in philosophy.

[1] Note that they are not always in conflict as Cooper and Abrams [15] point out. If the primary goal of a decision maker is to achieve fairness, then the first five value-laden questions of our framework are still relevant, but the sixth one is not of any interest.

Manuscript submitted to ACM

Philosophers have attempted to characterize ideally just institutions in terms of the set of moral principles they satisfy [53] or to characterize those, and only those, inequalities that are ultimately morally important for justice [62], thus providing a normative grounding to judgments about injustice reduction. So far, these two debates have mostly been developed apart from each other. However, the philosophical moral grounding problem is relevant for computer science since the criteria of fairness discussed within that discipline cannot be simultaneously fulfilled [8, 14, 40].

It is important to note that the debate about appropriate fairness metrics is not a mathematical debate [58, 64]. As Jacobs and Wallach [37] points out, the plethora of fairness definitions and the conflicts between them stem from the conflicting theories of fairness that they operationalize and reflect different values. Thus, it is a debate about values [37] and one's beliefs about the world [23]. Recent works [26, 64] have highlighted the need for a deliberative process to explicate these values. Wong [64] argues that the choice of fairness metric(s) is a choice of values and thereby inherently political. Consequently, [64] demands a democratization of this choice.

Here we propose a framework for eliciting and implementing moral values relevant to the choice of a fairness goal achievable by prediction-based decision-making. Our proposed framework elicits these values from decision makers and decision subjects through six value-laden questions. It also provides a simple way to set parameters of a prediction-based decision-making system such that it aligns with the agreed-on values. We assume a binary decision-making system where individuals are assigned probabilities, e.g., the predicted probability of repaying a loan. A decision rule takes this predicted probability as an input and makes the final binary decision. We also assume that it is possible to compare the consequences of these decisions for two socio-demographic groups (a privileged one and a disadvantaged one) in terms of the utility they generate for the decision subjects.

The central idea of the framework is to specify one's normative preferences regarding six value-laden questions:

(1) **Utility of the decision maker:** How should we assess the benefit/harm that the decision maker derives from the decisions?
(2) **Utility of the decision subjects:** How should we assess the benefit/harm that the decision subjects derive from the decisions?
(3) **Relevant groups:** What groups of people are affected unequally by decision-making systems because being a member of a group is a (direct or indirect) cause of inequality?
(4) **Claim differentiator:** By virtue of which features can individuals morally demand equal consideration by the decision maker?
(5) **Pattern of justice:** Should the goal of justice be equality or some other distribution (e.g., maximizing the expectations of the worst-off group)?
(6) **Trade-off decision:** How strongly should fairness be pursued if it comes into conflict with the utility of the decision maker?

As can be seen in Figure 1, question (1) helps to derive a score of the decision maker's utility. Questions (2)-(5) allow us to define a morally appropriate fairness criterion and a score that expresses to what degree it is fulfilled. Question (6) then balances these two scores through a Pareto front that compares different possible decision rules. This step thus makes the trade-off explicit.

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 perspectives: the decision maker's and the decision subjects'. In this context, we explain the first value-laden choice of our framework. In addition, we introduce the notation that we will use throughout the paper. In Section 4, we describe a common structure of

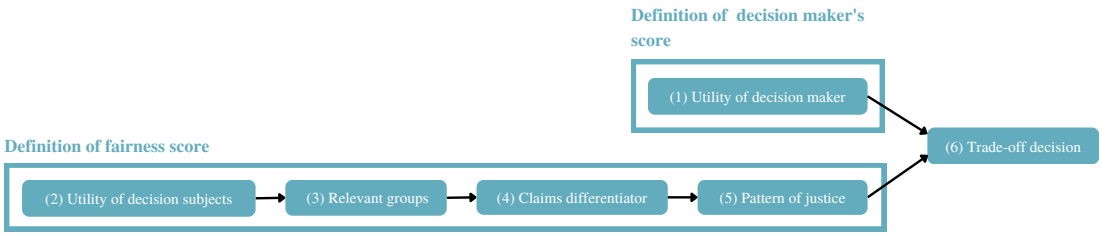


Fig. 1. The six steps of our framework and their connections.

theories of justice that is represented in steps (2)-(5) of our framework. In Section 5, we will present our suggestion for navigating the trade-off between the decision maker's goals and the fairness towards the decision subjects based on [39] — the final step of our framework. We will then exemplify the value-laden choices in a case study in Section 6. Finally, we discuss the limitations and merits of our framework in Section 7 and conclude in Section 8.

2 RELATED WORK

Our paper combines several distinct developments in the recent literature on fairness in machine learning. The goal is to use them in a novel way to define an ethical framework for supporting the implementation of a fairer decision-making system.

Utility-based view of fairness. The first development is framing the problem of fairness in machine learning as the moral problem of justifying the distributive implications of the decisions based on those predictions. As not every individual derives the same benefit or harm from the same decision, a line of research has developed that formalizes the *utility* or *welfare* implications of such decisions for the individuals affected by them (in our framework, "decision subjects"). As pointed out by philosophers in the debate on whether the metric of justice should be resources, utility, or capabilities [45, 59], a resource (e.g., a loan offered by a bank) can be converted into different utility or capability levels by different types of individuals (e.g., those who are able vs. not able to repay a loan).

In the history of fair ML debate, this approach has been pioneered by several papers at the intersection of economics, political philosophy, and machine learning. Heidari et al. [31] have proposed welfare-based definitions of fairness that take the effects of decisions into account and can be used as learning constraints. Heidari et al. [32] have highlighted that what a fair distribution of utility looks like is influenced by one's claim to utility — by which we mean the moral consideration that counts as a justification of inequality. In mapping the philosophical theory of equality of opportunity to group fairness metrics, they consider individual effort to be relevant for the moral claim one has to utility. This argument has been further developed in [10, 43]. Hertweck et al. [33] and Hu and Chen [36] showed, through philosophical arguments and with an empirical study (respectively), that enforcing fairness criteria can actually harm marginalized groups if it is not guided by utility considerations.

Drawing on this literature, our framework views fairness in machine learning as a special case of the problem of selecting a distributive mechanism that is reasonably expected to achieve justice. In this view, the algorithm plays an analogous role as social institutions in traditional theories of justice, as proposed by Heidari et al. [32].

Choice between conflicting fairness criteria. The second development in the literature is the emergence of multiple, conflicting fairness criteria, lacking a systematic framework to choose among them. These criteria can be categorized in different ways. Our approach can be seen as an extension of previous attempts to systematize the choice between *group fairness criteria*. Standard group fairness criteria compare metrics derived from the confusion matrix (such as true positive rates, false positive rates, positive predictive value, etc.) for two or more socio-demographic groups. Commonly, group fairness criteria demand equality in this comparison metric [12]. Saleiro et al. [57] and Makhlouf et al. [46] provide a flow chart to guide the choice between these standard fairness criteria.

We share the goal of these papers: to create a framework for — or even better, a morally principled solution to — the problem of selecting a pertinent fairness criterion. In some respect, our approach can be regarded as an extension (and generalization) of Heidari et al. [32], who 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. The novelty of our approach is to redefine the scope of the question: our problem is no longer to select between the items of an already given list of group fairness criteria (derived from the confusion matrix) but to formalize stakeholder values related to justice in such a way that a pertinent group fairness criterion will be determined. The group fairness criterion selected in this way can, but does not have to, correspond to one of those that can be specified by reference to the confusion matrix.[2]

Relation to individual fairness and causality-based fairness. Group fairness criteria are traditionally seen as opposed to individual fairness [19] or causal definitions of fairness [41].[3] We note, *inter alia*, that our paper addresses the two standard objections against group fairness criteria raised by these two alternative approaches: fairness gerrymandering [19] and causal irrelevance [41]. Philosophers Hedden [30] and Long [44] argued against classification parity (a specific statistical group fairness measure) by pointing to examples in which its violation was morally irrelevant. Remarkably, in both of these examples, the group variable is causally irrelevant by construction. Thus, these specific arguments against classification parity provide further support to the (much broader) causal irrelevance objection. We shall later discuss how our approach responds to the criticism in question.

Fairness-utility trade-offs. Finally, our framework brings together the aforementioned debate with proposals to evaluate trade-offs between the goal of the decision maker and fairness [39]. One option to balance these two goals is to train a model on an objective function that combines the decision maker's utility and a fairness score as seen in [38]. However, this option requires weighing the importance of both perspectives which can be difficult without having a clear idea of how big the trade-off may be. Several works (e.g., [9, 16, 27, 47]) have highlighted these conflicts and tried to quantify the trade-off between the two goals. However, these works demonstrate the trade-offs for specific instantiations of the decision maker's goal (such as accuracy) and fairness (standard group fairness criteria). 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 [39] 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. The framework we propose provides a simple approach that builds on [39] to support the question of how to balance this definition of fairness with the utility of the decision maker.

[2]This is explained in detail in [4], where it is also shown that standard group fairness metrics — i.e., (conditional) statistical parity, equalized odds, equality of opportunity, FPR parity, sufficiency, predictive parity, and FOR parity — can be derived for specific choices of steps (2)-(5) of our framework.
[3]However, see [12] for the claim that fair ML categories represent such distinctions as being sharper than they actually are.

Research gap. What is still missing is a unified framework that guides stakeholders to identify the type of fairness goal they want to achieve (assisted by a menu of standard normative choices from political philosophy) and the degree to which they want to achieve the goal in conflict with the main purpose of the prediction algorithm.

3 PREDICTION-BASED DECISION-MAKING

Our approach models fairness for prediction-based decision-making systems with distributive implications (i.e., purely predictive mechanisms relevant to epistemic views of justice [2, 22] fall outside the scope of this approach). The goal of these systems is to make a decision D based on a set of variables. Predictions are needed because the central variable that 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 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. -Here we shall assume that the decision maker pursues reasonable goals and that there exists a metric that expresses the degree to which these goals have been achieved. We refer to this metric as the "decision maker's utility". In prediction-based decision-making systems, the decision maker typically makes many decisions of the same type. Their possible consequences can be identified and modeled probabilistically. Thus, the degree to which the decision maker's goal is achieved can be measured as *expected utility* (utility weighted by probability). We assume that utility in this sense is something that decision-makers typically want to maximize.[4] This requires a **first value-laden choice**: how does one represent and measure the utility the decision maker wants to achieve through a given set of decisions?[5]

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 goal (which is measured by the utility function). This requires assessing 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 FAIRNESS METRICS FOR DECISION-MAKING

While attempting to achieve the goals for the decision maker, any prediction-based decision system relevant to our analysis coincidentally (and in some cases, unintentionally) distributes benefits (or harms) among members of society. We understand a fair prediction-based decision system to involve predictions and decision rules that combined can be reasonably expected to achieve a just distribution of benefits and harms across different groups. We turn to theories of justice in the tradition of political philosophy in order to determine what is a *just* distribution.

[4]Depending on the context, there might be different boundary conditions, such as resource constraints, legal obligations, or business strategies.
[5]The reason why we call this choice "value-laden" is that often it is impossible to derive a proper measure of utility in the sense we specified by simply observing the behaviors of decision-makers. In particular in complex organizations, morally significant choices (such as in human resources) often pursue several goals simultaneously. The definition of a goal (even when the goal is defined as a weighted function of a plurality of goals) always involves a drastic simplification from the observed social reality, which can hardly be achieved without relying on some normative assumption.

We characterize theories of justice by their answers to the following questions, which represent the **next value-laden choices**: What is, ultimately, distributed? Between whom is it distributed? Which subgroups should be compared? And how should it be distributed? [53, 59].

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 harm, as the *utility of the decision subjects*. This builds on the line of welfare-based definitions of fairness described in Section 2. Utility can be defined in different ways. We define well-being as what people have reasons to desire — an "objective list" or "informed-desire" approach [17, 25] and delegate the choice of a measure of utility to the hypothetical stakeholders that would employ this framework to arrive at their favored definition. Negative utility can be defined as what people desire *not* to have.

Definition 4.1 (Decision subject utility). Decision subject utility is the amount of benefit or harm derived from receiving a certain decision. It is what people have (objective) reasons to desire.

In our framework, we do not consider the overall level of utility of decision subjects but only the utility that is gained or lost as a result of the decisions taken with the aid of the algorithm.

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. We rely on competent experts and stakeholders to identify a suitable operationalization of the concept of utility into something measurable, which is the second value-laden choice of our framework.

4.2 Relevant groups

Between whom is it distributed?

Most contemporary theories of justice focus on individuals, understood as bearers of utility, capabilities, or rights [59]. Theories of discrimination, instead, relate to socially salient groups [1]. We focus on a conception of "relevant groups", placing causal constraints on what qualifies *a group* in a way that is relevant to group fairness. In our framework, relevant groups are defined by a *weak causal link* in the context of the prediction-based decision in question.

Definition 4.2 (Relevant groups). Relevant groups are types of individuals that are representative of plausible causes of inequality in the outcome or in the prediction in the context to which the question of fairness relates.

By invoking groups that satisfy a weak causal link we aim to address the objections typically raised against group fairness that we already mentioned in Section 2, namely fairness gerrymandering and causal irrelevance. Proponents of individual fairness object that group fairness criteria are vulnerable to fairness gerrymandering: Any group fairness criterion can be satisfied by altering who receives a positive and negative decision, without improving the fairness of the treatment of any individual in that group [9, 19, 55]. Proponents of causal definitions of fairness object that group fairness criteria are unable to distinguish the case in which individuals of a group receive a worse outcome *because* they are members of the group from those cases in which receiving a worse outcome is simply *correlated* to being a member of the group, but the group does not as such *influence* the decision. Our weak causality requirement demands to only consider groups defined by features that are *plausible* causal influences of the prediction or the outcome (or both), where causation can be both direct and indirect. So, for example, in a racist society, race may define relevant

groups; in a sexist society, gender may define relevant groups. The Cartesian product of the two (each combination of a race and gender variable) will then also be weakly causally relevant. Unlike counterfactual fairness, which requires modeling causal links between the prediction variable and the group variable, our framework takes the shortcut of only considering groups defined by features for which some degree of causal influence on the decision-relevant variable or the decision is plausible *a-priori*, given what we know about society. In principle, these groups *could* be defined to be narrower and narrower. This corresponds to the concept of multicalibration, which considers every efficiently-identifiable subgroup, i.e., the "collection of subsets where set membership can be determined efficiently — for instance, subpopulations defined by the conjunctions of a small number of boolean features or by small decision trees" [29, p. 1940]. However, inequalities between very large numbers of extremely fine-grained groups are hard to morally judge in practice. Therefore, in contrast to multicalibration, we consider only groups for which the weak causality requirement is satisfied instead of constraining the number of considered subgroups purely based on computational efficiency. This means that our framework aims to compare groups (including those with very few and very similar individuals) identified by all the features that causally influence (directly or indirectly) the outcome or the prediction. In most concrete contexts, a value-laden choice must be made to focus on one, or a few (intersectional) traits, guided by the concrete political priorities emerging in the context of our decision-making system and regarded most relevant by the stakeholders (step 3 of our framework). Admittedly, this offers no guarantee that every observed inequality in average outcomes between groups is fully causally explained by the membership in those groups, so the approach is still vulnerable to counterexamples. However, our conjecture is that the causal requirement makes it harder, in practice, to gerrymander fairness by characterizing inclusion and exclusion criteria of groups in an arbitrary manner (just for the sake of equalizing group frequencies) and it will raise the chances that the observed group inequality is — to some degree if not entirely — due to the groups being what they are. We offer this as a mere empirical conjecture and as a pragmatic solution of the fairness gerrymandering and causal irrelevance problem. It is a "solution" in a very different sense than the rigorous (formal) solutions offered by individual fairness [19] and causal views of fairness [41]. As a practical method, our framework makes approximate fairness easier to achieve practice, because it has lower epistemic requirements than competing approaches, such as individual and counterfactual fairness.[6]

4.3 Claim differentiator

Which subgroups should be compared?

In answering the question, "between whom are benefits and harms distributed", we must consider the following complications. In some contexts, comparisons between groups (defined by causal relevance) are not intuitively appropriate for fairness. This is because, in some cases, individuals within those groups who are different in some (morally salient) features should not be treated equally. According to contextualism[48], what these morally salient features are depends on the context. For example, there are contexts in which individuals ought to be treated differently when their needs differ, but in other contexts individuals ought to be treated differently when their contributions differ. Moreover, in practice, we must deal with moral disagreement about whether need, responsibility, or contribution, for example, ought to matter, in a given context. To account for both contextual relativity and the possibility that not all stakeholder groups

[6] Specifically, individual fairness can only be measured relative to an already given metric of similarity of individuals in the respect that matters to fairness. Clearly, unless the metric is itself objectively fair in a morally relevant sense, individually fair predictors cannot be considered fairer in a substantive sense than the predictors they aim to improve upon. For this, see [12] showing that this is in all but artificially defined cases extremely hard to satisfy. Counterfactual fairness, on the other hand, can only be defined relative to an already given set of assumptions about the causal structure responsible for outcomes and decisions (e.g., a set of differential equations describing the direct and indirect influence of group features on both Y and D). In practice, it is extremely hard to know, justifiably believe, or even merely inter-subjectively agree upon a metric of similarity and a causal structure in the domains in which standard problems of fair AI emerge.

will adopt the same view of justified inequalities, we need to introduce a new parameter in the theory. This is the *claim differentiator*, the feature(s) by virtue of which individuals can morally demand equal consideration of the harms and benefits produced by the algorithms. In practice, this parameter specifies the subgroups among the previously defined *relevant groups* that should be compared. This is the fourth value-laden choice of our framework.

Definition 4.3 (Claim differentiator). A claim differentiator is a feature that distinguishes individuals who have different moral claims to utility.

We assume that the higher-order concept of a claim differentiator can be chosen on the basis of either the context or the moral theory endorsed by stakeholders. We introduce the claim differentiator as a novel concept that is not an established notion in political philosophy or moral philosophy.[7] To provide more clarity about this higher-order moral concept, we provide in Section 4.5.1 an analysis of luck-egalitarian equality of opportunity as the combination of substantive conception of the claim differentiator and a substantive conception of the pattern of justice, the element we introduce next. Moreover, we shall provide an example that illustrates the reasoning for the claim differentiator in a hypothetical concrete business scenario in Section 6.4.

4.4 Pattern of justice

How should it be distributed?

After discussing which groups have equal moral claims to the utility derived from the decisions, we have to consider 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., [11, 15, 36]). To avoid this, we can allow for some inequalities, e.g., if they are beneficial to the worst-off group. Therefore, the fifth value-laden choice is to define what a just distribution looks like. This can be described as a *pattern of justice*.

Definition 4.4 (Pattern of justice). A pattern of justice describes how utility should be distributed between the relevant groups.

The most widely discussed patterns of justice in political philosophy are:

- Egalitarianism [5]: The group utility levels should be as equal as possible.
- Maximin [53, 54]: The goal is to maximize utility for the worst-off group.
- Prioritarianism [35]: The goal is to maximize aggregate utility for all groups, giving greater weight to utility, the worse off the group.[8]
- Sufficientarianism [63]: The goal is to bring all groups above a certain level of utility.

This also implies that the patterns have a different relationship to equality. Egalitarianism values equality above all else while the other patterns tolerate inequalities: Maximin tolerates inequalities if they profit the worst-off group; prioritarianism tolerates inequalities if they increase the aggregated utility; sufficientarianism tolerates inequalities as long as all groups achieve a minimum level of utility.

[7] A *similar* idea is found in [10, 34, 43].
[8] Maximin is the extreme version of this as an infinite weight is given to the worst-off group.

4.5 Combining relevant groups, claim differentiators and pattern of justice

4.5.1 Combining relevant groups and claim differentiators. Both concepts of relevant groups and the claim differentiator split the general population into subgroups. We can connect the two concepts by asking: Should we compare the relevant groups or only subgroups of the relevant groups? This is relevant as the relevant groups are, of course, not homogeneous but consist of many different individuals who may have different claims to utility. Figure 2 shows this intuition of the claim differentiator as the selection of subgroups within the relevant groups. When we combine these two concepts, this gives us the groups whose utilities are to be compared from a fairness perspective: We analyze the distribution of utility between relevant groups, restricted to those individuals with the same moral claims (specified by the claim differentiator).

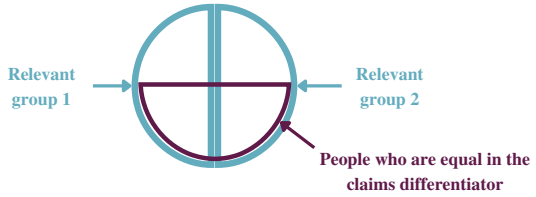


Fig. 2. The relationship of relevant groups and the claim differentiator.

We provide a concrete example of the combination of relevant groups and the claim differentiator in Sections 6.3 and 6.4.

4.5.2 Combining patterns of justice and claim differentiators. We can view many influential 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]. Luck-egalitarian equality of opportunity can be defined as the view that individuals who make similar choices should have the same expectations or outcomes [6]. This is different from strict egalitarianism, which demands equality in outcomes [42]. They differ in the claim differentiator: inequalities due to choices, not circumstances, are considered justified[7, 56]. Luck egalitarian equality of opportunity thus uses choices for which individuals are responsible as a claim differentiator.

It may be tempting to assume that the aforementioned definition of the claim differentiator implies that justice requires some kind of equality because individuals with the same value of the claim differentiator have morally the same claims to utility. However, this is wrong. Consider, for example, the combination of desert as a claim differentiator and maxmin as a pattern of justice. Justice is then achieved by maximizing the expectations of the worst-off relevant group *among individuals who are equal in their contributions*. This is not a morally absurd view. For example, one may object to achieving equality between equal contributors (while recognizing that this is what they ideally deserve) when this, in the given circumstances, can only be achieved by leveling down.

5 TRADE-OFF BETWEEN THE DECISION MAKER'S GOALS AND FAIRNESS

The previous steps allow us to define a fairness score, which quantifies the fairness of a decision rule, in a way that encapsulates the value choices listed above.[9] Given a fairness score and a measure of expected decision-maker utility, it is possible to represent trade-offs in a bidimensional Cartesian plot [39]. It is reasonable to focus on the Pareto-efficient

[9]The mathematical details of how exactly to derive a fairness score from these fairness components is described in full detail in [4].

decision rules: those for which an improvement on one dimension is only possible if the other dimension is worsened. The last stage of moral discussion ought to concern the choice between points in the Pareto front, where any gain of fairness can only be achieved at the expense of the decision maker's utility, and vice-versa.

6 CREDIT LENDING EXAMPLE

Let us now discuss a highly simplified example of financial lending to see how the perspectives of the decision maker (Section 3) and the decision subjects (Section 4) can be defined and balanced in practice. Consider a bank's decision to accept a loan application ($D = 1$) or reject it ($D = 0$) based on scores representing repayment probabilities. If the bank wants to balance its profitability with fairness, the first step is to specify how they measure its own utility, i.e., its profits. The subsequent steps of our framework determine how they measure fairness. To illustrate this example, we are using the preprocessed version of the UCI German credit dataset and train a logistic regression to predict whether an individual will pay back a granted loan [24].[10]

6.1 Utility of the decision maker

We start by specifying the perspective of the decision maker. Assuming that the bank is interested in profits, it has to assess how much profit is derived from each decision. As this illustration does not aim to be realistic, we do not consider the costs of the bank and assume that the interest payments[11] are the profit of the bank while the cost of a defaulted loan is equivalent to the loan size. Rejected loan applications are defined as cost-neutral as we assume that the cost of reviewing applications is 0. 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 thus $E(u_{DM,i}) = p_i \cdot z_i \cdot s_i - (1 - p_i) \cdot s_i$. Utility-maximizing banks would grant a loan to all individuals with a positive expected utility (i.e., $D = 1$ if $E(u_{DM,i}) > 0$).

6.2 Utility of the decision subjects

Next, the bank turns to the evaluation of how fair a given decision-making system is towards the decision subjects. For this, they might ask representatives of the decision subjects or experts from social sciences and philosophy to consider the components of fair decision-making described in Section 4.

These representatives first have to answer the question of how to assess the utility that decision subjects derive from the decision-making process. In the case of lending, *loans* are distributed. Individuals do not profit equally from being granted a loan. If they cannot repay the loan and end up defaulting, it harms their future chances of receiving a loan. To keep this example simple,[12] we assume that stakeholders will base the utility assessment on the decision D and whether the individual repays the granted loan Y. In our case, D and Y are binary variables, so there are four combinations whose utility we have to determine.

The utilities of the decision subjects can be visualized as a 2x2 matrix. Utility weights can be elicited through a dialogue between relevant stakeholders and experts on the impact of financial decisions. We note that, as long as there is a well-specified reference point and a corresponding scaling factor [20], it is not necessary to express utility weights in terms of an external dimension (e.g., money) - weights matter only in so far as they define how morally beneficial or harmful a consequence is *in proportion* to a different one. For example:

[10] The code for this has been attached to the submission and will be made publicly available after the acceptance of the manuscript.
[11] In the implementation of this example, we assume the interest rate to be 10% for every loan regardless of risk or other factors.
[12] Additionally, we may consider factors such as the loan size 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).

- $u_{D=1,Y=1}$: 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 assign a utility of +10.
- $u_{D=1,Y=0}$: 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.
- $u_{D=0,Y=1}$: 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.
- $u_{D=0,Y=0}$: 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.

If the property $u_{D=0,Y=0} \neq u_{D=1,Y=1}$ holds, we can to fix $u'_{D=0,Y=0} = 0$ and $u'_{D=1,Y=1} = 1$ so that the remaining utility weights can be expressed relative to this reference point and scale: $u'_{D=1,Y=0} = (u_{D=1,Y=0} - u_{D=0,Y=0})/(u_{D=1,Y=1} - u_{D=0,Y=0})$ and $u'_{D=0,Y=1} = (u_{D=0,Y=1} - u_{D=0,Y=0})/(u_{D=1,Y=1} - u_{D=0,Y=0})$. This results in the utility matrix visualized in Table 1. Notice that this shifting and scaling of all entries of the utility matrix does not affect the final decisions. In practice, it is crucial to always elicit utility weights relative to a well-specified baseline.

Table 1. 2x2 matrix representing the utility of the decision subjects

	Y=0	Y=1
D=0	0	-1
D=1	-5	+10

6.3 Relevant groups

The representatives next have to define the relevant groups to compare and agree that groups defined by the sex attribute have unjustly unequal chances in life.[13] They therefore determine that the relevant groups to compare are women and men.

6.4 Claim differentiator

To determine the claim differentiator, the representatives have to answer the question "What makes it the case that *certain individual types* (groups of people) have roughly the same claims to utility?" Suppose that the representatives agree that loan defaulters and non-defaulters cannot demand equal consideration. The only clients who have a claim to benefit from the decisions are those who will repay their debt. Therefore, they will compare the utility of people who repay their loan ($Y = 1$).

[13] Even though sex is not binary, it is represented as a binary variable in this dataset.

6.5 Pattern of justice

We suppose that after deliberation, the representatives agree on the maximin pattern, so that the fairness score increases as the utility of the worst-off group increases. When the bank and representatives compare different decision rules, they have to analyze how well these decision rules do with respect to maximin among the non-defaulters. This requires computing the expected utilities for both male and female non-defaulters under each decision rule and then comparing the lowest expected utilities.

6.6 Trade-off decision

From steps (2) to (5), it follows that the representatives decided to maximize the utility of the worst-off group between women and men who repay their loans ($Y = 1$). However, suppose that this fairness goal conflicts with the decision maker's utility defined in step (1). The last step in our framework is therefore to look at the trade-off between the goals of the decision maker and the fairness towards decision subjects. As described in Section 5, we use a Pareto front to visualize this trade-off for many different decision rules. In line with [9, 16], we will assume that the decision rule takes the form of a threshold.[14]

In this example, we test upper- and lower-bound thresholds for each group (men and women), resulting in the $(2 * 101)^2$ points seen in the Pareto plot in Figure 5 (in the Appendix), where the Pareto front is marked in blue.[15] The y-axis shows the decision maker's average utility per customer. A utility of, e.g., 10, means that the bank can expect a utility of 10 Deutsche Mark per customer. The x-axis shows the fairness score, which is the lower utility between the utility of women with $Y = 1$ and men with $Y = 1$.

Of course, we cannot claim that this Pareto plot shows the entire Pareto frontier as more points could be added. However, it visualizes some representative elements in the trade-off.

As a start, the stakeholders, i.e., the bank and the representatives, may 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 34). As can be seen in Figure 4, maximizing the utility of the decision maker results leads to inequality in the utility of women and men where women have the lower utility. With increasing fairness, the utility of the worst-off group (marked in yellow) also continuously increases. However, the difference in women's and men's utility does not continuously decrease even though the average utilities end up converging to the maximum possible expected value of 10 (which is achieved when everyone who is able to repay their loan receives a loan) for the fairest point (point 34).

Figures 5 (in the Appendix) and 4 also show the other points on the Pareto front and the corresponding utility values for women and men. As can be seen, most points on this Pareto front would lead to a negative expected utility for the bank (points below the red line in Figures 3 and 5). The bank does of course not consider such decision rules as it would sooner or later go out of business. Figure 3 therefore focuses on the profitable decision rules. Among those, the stakeholders may see points 3-5 as good trade-offs: They achieve 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 (point 2), so the representatives may argue that this gain in the utility of the decision maker is too costly in terms of fairness.

It is important to note that our framework offers no principled solution to the problem of determining a valid trade-off value. The last step, unlike the previous five, is not guided by any kind of theory but it expresses the actual degree

[14] [9, 16] have proven this for egalitarian fairness criteria. We leave the proof that this also holds for maximin to future work.
[15] 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.

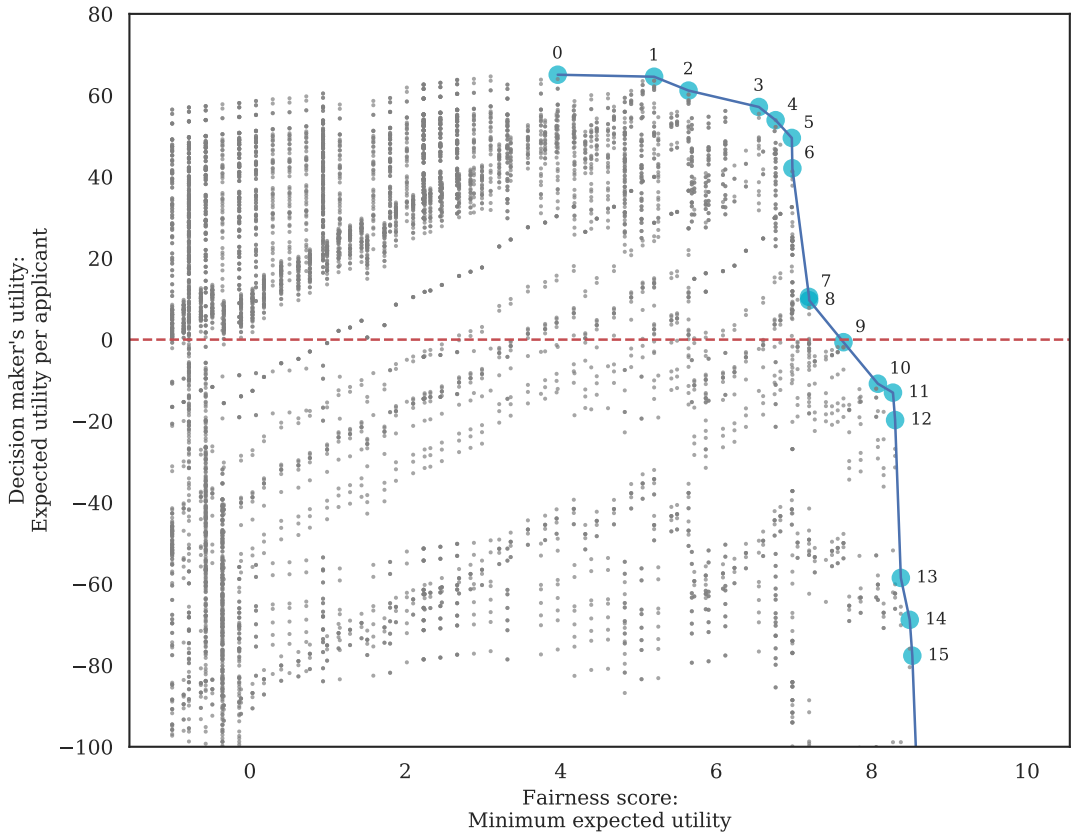


Fig. 3. The Pareto front where the small, gray points are Pareto-dominated by the larger, blue points. Zoomed in to focus on the decision rules that are profitable for the bank.

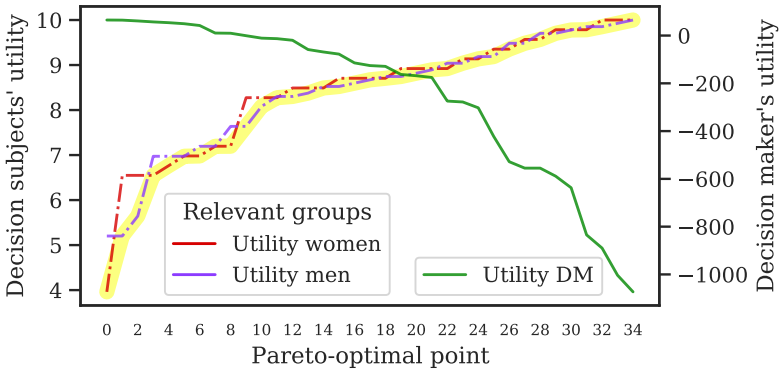


Fig. 4. The DS utilities of women and men resulting from the decision rules on the Pareto front. The minimum expected DS utility (i.e., the fairness score for maximin) is highlighted in yellow.

of attraction to fairness and aversion to loss of utility for the stakeholders consulted for this decision. The value of the framework is to lead the stakeholders to discuss *necessary* sacrifices of utility or fairness, avoiding the selection of decision rules that are not Pareto optimal (given the value assumptions that have been - we suppose - antecedently agreed upon).

7 LIMITATIONS

Interpersonal comparisons of well-being are notoriously difficult and here we rely on an objectivist view of well-being (which has to be elicited from experts) that may not correspond to the preferences and beliefs of the people affected by the decisions. Furthermore, this approach is welfarist throughout. Welfarism is criticized by proponents of the capability approach as being too subjective [60]. Notice, however, that we do not rely on a preference-based or pleasured-based account of well-being, so we can include benefits to individual autonomy and freedom into our utility metric if stakeholders can agree on a suitable measurement. An objective measure of well-being can also consider the impact of resources on real individual freedoms (also known as capabilities) [45].[16]

Moreover, 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 [51].

On a practical level, it is not obvious how to make the six value-laden choices in practice and we only provided a sketch for this. More work needs to be done to deliver a practical empirical methodology to elicit the relevant value-laden choices from stakeholders.

This is perhaps why current group fairness metrics are so tempting: They do not require us to think through the choices of our framework. However, we must not delude ourselves: Not specifying every value-laden choice in our framework does not mean that we remain agnostic about what an appropriate choice might be — we simply choose it implicitly. We argue that it is preferable to make these value-laden choices explicit in the design process.

8 CONCLUSION

With the increasing use of automated decision-making, there is a rising need to develop these systems ethically. This is not just a technical question, but a question of values. However, these values are typically hidden in the technical details of the implementation. Like others before us, we therefore advocate for a more public debate about the values implemented in decision-making systems.

In this paper, we offer a framework to reveal and specify six key value-laden decisions behind the implementation of prediction-based decision-making systems. This includes the choice of a fairness criterion and the degree to which it can be enforced compatibly with the decision makers' original intentions.

Our framework helps bring the debate about values to the forefront of implementation, rather than leaving these values as an accidental by-product. While our framework models more complex moral options than most, we kept it simple enough to be usable for actual stakeholder debates. We developed a web application that supports this deliberation process by visualizing fairness scores and their relation to the decision maker's utility.[17]

[16] In this regard, we register disagreement between those, like Sen [61], who maintain that the value of individual agency cannot be captured by a welfarist framework and those, like Griffin [25] who maintain that it is an element of (objective) well-being.
[17] Link to web app removed for anonymous review.

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] Elizabeth Anderson. 2012. Epistemic justice as a virtue of social institutions. *Social epistemology* 26, 2 (2012), 163–173.

[3] 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

[4] Anonymous. 2023. Distributive Justice as the Foundational Premise of Fair ML: Unification, Interpretation, and Extension of Group Fairness Metrics. (2023). Unpublished manuscript – submitted as part of the supplementary material.

[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] Richard J. Arneson. 1989. Equality and Equal Opportunity for Welfare. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 56, 1 (1989), 77–93.

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

[9] 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.

[10] 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 (SDS)*. 19–25. https://doi.org/10.1109/SDS54800.2022.00011

[11] 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

[12] Reuben Binns. 2020. On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 514–524.

[13] 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.

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

[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] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

[20] Charles Elkan. 2001. The Foundations of Cost-Sensitive Learning. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2 (IJCAI'01)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 973–978.

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

[22] Miranda Fricker. 2007. *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press.

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

[24] 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.

[25] James Griffin. 1986. *Well-Being: Its Meaning, Measurement, and Moral Importance.* Clarendon Press.

[26] 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.

[27] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016).

[28] 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/

[29] Ursula Hebert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. 2018. Multicalibration: Calibration for the (Computationally-Identifiable) Masses. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer Dy and Andreas Krause (Eds.). PMLR, 1939–1948. https://proceedings.mlr.press/v80/hebert-johnson18a.html

[30] Brian Hedden. 2021. On statistical criteria of algorithmic fairness. *Philosophy and Public Affairs* 49, 2 (2021).

[31] 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).

[32] 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.

[33] 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

[34] Sune Holm. 2022. The Fairness in Algorithmic Fairness. *Res Publica* (2022). https://doi.org/10.1007/s11158-022-09546-3

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

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

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

[38] Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. 2011. Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*. IEEE, 643–650.

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

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

[41] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. *Advances in neural information processing systems* 30 (2017).

[42] 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.

[43] 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.

[44] Robert Long. 2021. Fairness in machine learning: Against false positive rate equality as a measure of fairness. *Journal of Moral Philosophy* 19, 1 (2021), 49–78.

[45] Alan Lundgard. 2020. Measuring justice in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 680–680.

[46] 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

[47] 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.

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

[49] 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.

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

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

[52] 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.

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

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

[55] Tim Räz. 2021. Group Fairness: Independence Revisited. 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, 129–137. https://doi.org/10.1145/3442188.3445876

[56] John E Roemer. 1998. *Equality of Opportunity*. Harvard University Press, Cambridge, Massachusetts.

[57] 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).

[58] 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.

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

[60] 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

[61] Amartya Sen. 1997. *Choice, welfare and measurement.* Harvard University Press.
[62] Amartya Sen. 2009. *The Idea of Justice.* Harvard University Press, Cambridge, Massachusetts.
[63] Liam Shields. 2020. Sufficientarianism. *Philosophy Compass* 15, 11 (2020), e12704. https://doi.org/10.1111/phc3.12704
[64] Pak-Hang Wong. 2020. Democratizing algorithmic fairness. *Philosophy & Technology* 33, 2 (2020), 225–244.

A ADDITIONAL VISUALIZATION

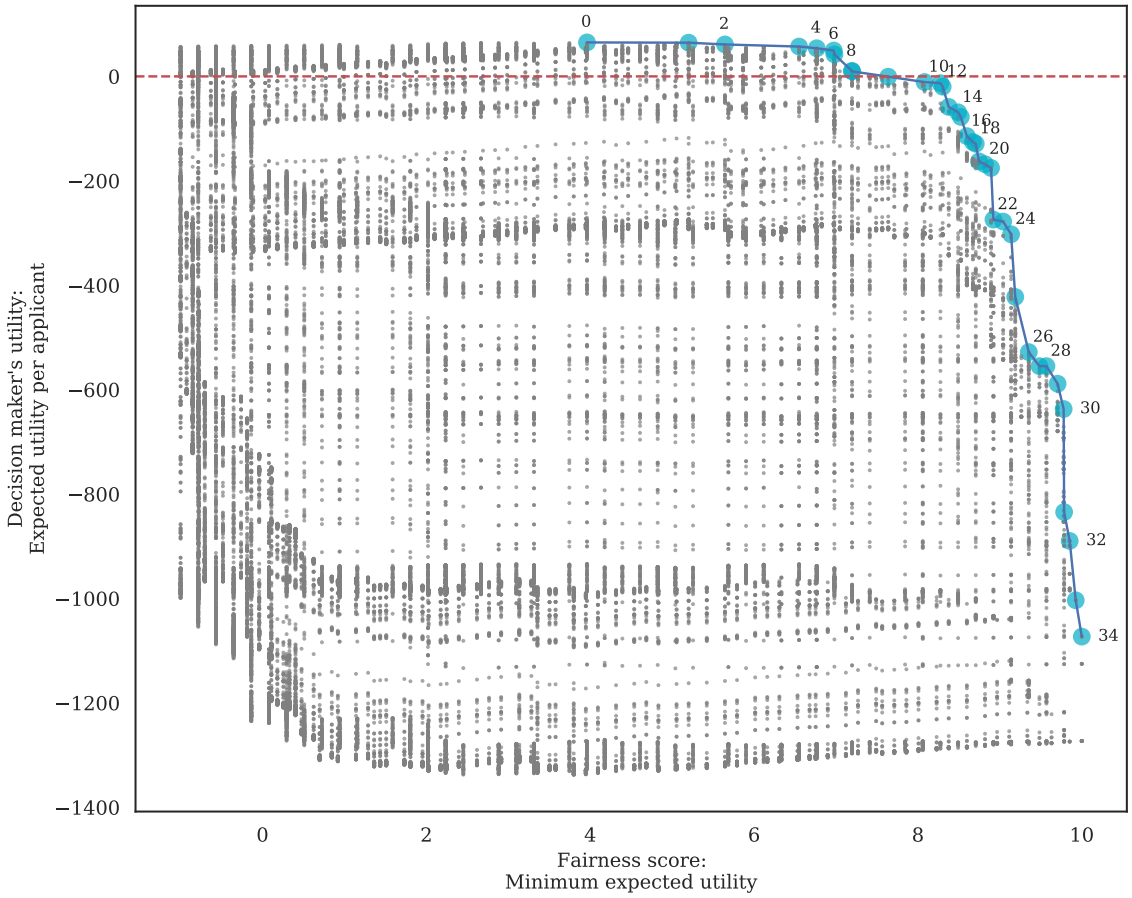


Fig. 5. The full Pareto front where the small, gray points are Pareto-dominated by the larger, blue points.