What's Distributive Justice Got to Do with It? Rethinking Algorithmic Fairness from the Perspective of Approximate Justice

Corinna Hertweck,[1, 2] **Christoph Heitz,**[1] **Michele Loi**[3]

[1]Institute for Data Analysis and Process Design, Zurich University of Applied Sciences, Winterthur, Switzerland
[2]Department of Informatics, University of Zurich, Zurich, Switzerland
[3]AlgorithmWatch, Berlin, Germany
corinna.hertweck@uzh.ch, christoph.heitz@zhaw.ch, loi@algorithmwatch.org

Abstract

In the field of algorithmic fairness, many fairness criteria have been proposed. Oftentimes, their proposal is only accompanied by a loose link to ideas from moral philosophy – which makes it difficult to understand when the proposed criteria should be used to evaluate the fairness of a decision-making system. More recently, researchers have thus retroactively tried to tie existing fairness criteria to philosophical concepts. Group fairness criteria have typically been linked to egalitarianism, a theory of distributive justice. This makes it tempting to believe that fairness criteria mathematically represent ideals of distributive justice and this is indeed how they are typically portrayed. In this paper, we will discuss why the current approach of linking algorithmic fairness and distributive justice is too simplistic and, hence, insufficient. We argue that in the context of imperfect decision-making systems – which is what we deal with in algorithmic fairness – we should not only care about what the ideal distribution of benefits/harms among individuals would look like but also about how deviations from said ideal are distributed. Our claim is that algorithmic fairness is concerned with unfairness in these deviations. This requires us to rethink the way in which we, as algorithmic fairness researchers, view distributive justice and use fairness criteria.

1 Introduction

In the algorithmic fairness literature, there are numerous fairness criteria and metrics that try to operationalize the complex and context-dependent construct that is fairness (Jacobs and Wallach 2021). As these criteria and metrics have often been proposed without a deeper engagement with the philosophical literature, there is a wave of work that has tried to tie the proposed statistical metrics of fairness to the philosophical debate. This is particularly the case for group fairness criteria and egalitarianism (see, e.g., Heidari et al. (2019); Binns (2018, 2020); Gajane and Pechenizkiy (2017)). Egalitarianism is a theory of distributive justice and, as such, tells us something about how benefits/harms should be distributed among members of a society. There is thus a parallel to algorithmic decision-making systems that make decisions about individuals that result in different benefits/harms and could therefore be seen as distributing benefits/harms in a population. The fact that fairness criteria for decision-making systems have been linked to distributive justice thus seems intuitive. However, we believe that this often-made connection is blurred and leaves an important conceptual question unanswered: *If theories of distributive justice are defined at the level of individuals, how do they relate to group fairness criteria defined at the level of socio-demographic groups?* In this paper, we use this question as a starting point to understand the relationship between distributive justice and algorithmic fairness. Based on this, we present our understanding of what we think is usually meant by "algorithmic fairness". The crux is this: Patterns of distributive justice like egalitarianism, sufficientarianism or maximin cannot be perfectly fulfilled in practice. If a pattern was perfectly fulfilled (e.g., for egalitarianism: all individuals receive the same outcome), then there would be no structural injustice (in how the pattern is fulfilled). If, however, the pattern is not perfectly fulfilled – which is the case with any real-world system as there will always be deviations from the ideal – then individual deviations from the ideal distribution (i.e., when an individual does not receive what they should receive under the ideal distribution)[1] might affect some groups more than others – in a biased way. This is what we will call *structural injustices*.[2] Generally speak-

[1]There could be multiple ideal distributions. There are, for example, multiple possible distributions that fulfill egalitarianism. One then has to decide how to measure the deviation from the ideal by deciding on one ideal. For egalitarianism, one option would be to measure the deviations from the distribution where everyone gets the mean of what everyone gets under the current distribution.

[2]Note that our use of the term "structural injustices" is not identical to Iris Marion Young Young (2010)'s use of the term. Kasirzadeh (2022) already discusses the relationship between structural injustice as defined by Iris Marion Young and algorithmic fairness (which is more similar to our definition of the term). She finds that "algorithmic fairness does not accommodate structural injustices in its current scope" (Kasirzadeh 2022, p. 349) because mathematical fairness criteria do not consider the sociotechnical nature of automated decision-making systems and instead just condense fairness into a single metric that does not consider "power structures and social dynamics" (Kasirzadeh 2022, p. 351). We strongly agree with this view and urge practitioners not to misinterpret the fulfillment of statistical criteria as proof of justice or fairness. However, our paper will interpret the term "structural injustice" only in the context of what we can evaluate given the data

ing, structural injustices are, for example, unfair opportunities (which are unjust according to Rawls (1999) fair equality of opportunity principle) or systematic, structural discrimination against a certain category of people, which is also clearly unjust. In the terminology of distributions, we measure structural injustices as unfairly distributed individual deviations from an ideal distribution. We argue that this is what algorithmic fairness metrics are supposed to capture even though the algorithmic fairness literature has not made this clear so far. Note that this is *our* interpretation of the fairness concern in the algorithmic fairness literature because there is no agreed-upon definition of what is meant by algorithmic fairness. Our approach is different from others as algorithmic fairness is typically associated with theories of distributive justice themselves. We, however, argue that algorithmic fairness metrics do **not** check whether a theory of distributive justice is fulfilled, and instead, it is the deviation from said theory that fairness is concerned with. We thus rethink every theory of distributive justice as follows:[3] There is a distributive pattern that defines how benefits/harms should be distributed between individuals. If this is perfectly fulfilled, then this is where the evaluation of the fulfillment of the theory of distributive justice ends. However, if the distribution system makes mistakes, then it is not enough to test for the pattern of justice and we also need to test for structural injustice.

In Section 2, we will retrace how algorithmic fairness, and more specifically, group fairness, has been connected to theories of distributive justice thus far. Based on the issues we find with this, we present our proposal for how to map algorithmic fairness to distributive justice in Section 3. Next, we demonstrate what this means for fairness criteria in the context of distributions that are supposed to follow *egalitarianism*, *sufficientarianism* or *maximin* (Section 4). Section 5 compares the resulting fairness criteria from Section 4 to more naive fairness criteria that are based on the current status of the debate. Section 6 discusses how the work of Mittelstadt, Wachter, and Russell (2023) fits into our conceptual understanding of algorithmic fairness. Finally, we discuss the implications of our work and possible future directions of research in Section 7.

2 Retracing the Debate: Distributive Justice and Group Fairness Criteria

The parallels between distributive justice and automated decision-making systems mentioned in Section 1 have led to the algorithmic fairness literature drawing parallels between distributive justice and algorithmic fairness. This section will retrace this development after introducing the basics of both concepts.

of a decision-making system (and we will therefore sidestep questions of, e.g., procedural fairness).

[3]Note that we will assume that they are defined at the individual level, which is the case for the majority of them. An exception to this is Rawls's theory of justice (Rawls 1999).

2.1 Theories of Distributive Justice

Theories of distributive justice propose how benefits/burdens should be distributed in society and can vary in whether they care about the just distribution between individuals or groups (Lamont and Favor 2017). However, theories of distributive justice are typically concerned with individuals (Sen 1980). One notable exception from this is Rawls' difference principle, which is always illustrated by considering the average expectations of different "social classes" (Rawls 1999, p. 67) and what Rawls refers to as the "representative man [sic.] who is worst off" (Rawls 1999, p. 68). This representative person is a clear statistical or sociological generalization. However, the subsequent discussion within analytic philosophy focuses on a more abstract conception of justice disconnected from sociological categories. This is important to note as this makes it difficult to think about structural injustice at the group level for such theories.

Egalitarianism and Beyond The idea of **egalitarianism** is that equality per se is morally desirable. According to egalitarianism, people should thus receive the same outcomes, be treated in the same way, be seen in the same way or be equal in some other way. Egalitarianism can thus take different forms depending on what one sees as the metric that ought to be equal.[4] While egalitarianism is intuitive, it has one remarkably unpalatable implication. Egalitarianism between two people can be achieved not only by increasing the benefit of one person but also by lowering the benefit of the other person (or even by lowering the benefit of both to a lower level). Thus, egalitarianism can be achieved by worsening the results for both people, which is often referred to as the "leveling down" objection. A reasonable agent may reflect that there are contexts in which achieving equality by leveling down would be considered a worsening from the point of view of fairness by most affected people. It would then seem unreasonable to impose fairness that may worsen the prospects of all the groups involved – especially when the affected groups themselves refuse it.

Thus, if one cares about the social acceptability of fair prediction-based decisions, one should either abandon egalitarianism or suitably modify it in order to align it with stakeholders' view of fairness – at least to the extent that these appear to be reasonable. The most obvious way to do this is suggested by the history of the debate on the nature of justice in analytic political philosophy. Since the seminal work of John Rawls (1999), alternatives to simple equality have

[4]Note that egalitarians advocate for equality in a qualified form. A common form of egalitarianism is, for example, "luck egalitarianism": Both Cohen (1990) and Temkin (2017) defend this idea of equality (even at the cost of leveling down) as the best description of what justice is (before compromise with other values). What they mean by equality is "equality in conditions that are not the result of choice". A plausible alternative is to say that equality is an appealing idea when it is framed in terms of equal desert. This is the "desert egalitarian" form. Kagan (2012) frames "desert" as moral desert. Alternatively, desert can also be a placeholder for any other justification of inequality (Loi, Herlitz, and Heidari 2023), e.g., when used in generic sentences such as "people with more urgent needs deserve more urgent assistance by the state".

emerged, such as the **difference principle**, which requires maximizing the resources of the least advantaged group in absolute terms. Note that the difference principle is the application of maxmin to a particular context and given specific boundary conditions. Hence, we refer to **maxmin** in the context of various operationalizations where the context differs (sometimes significantly, sometimes in subtle ways) from the one of Rawls's difference principle. **Sufficientarianism** also does not recognize equality's intrinsic value: According to sufficientarianism, equality simply has no moral value once all people's needs are satisfied, and it only has value to the extent that it is instrumental in meeting people's needs, or some other threshold of well-being that is equally significant for a human person as having a fundamental need fulfilled (e.g., a "non-basic" need, such as personal realization) (Frankfurt 2018). This alternative becomes particularly persuasive when the quantity of a limited resource deemed sufficient aligns with the quantity necessary to prevent significant harm or catastrophe. Consider, for example, a scenario where there is a limited supply of a particular resource, such as food or medicine, which is sufficient to sustain some but not all individuals within a population. Let us assume the population consists of ten members, each requiring a minimum of five units of the resource to survive, with a total of forty units available. In such a case, for any members of the population to survive, some must receive a larger share than others. Distributing the resource equally, providing each person with four units, results in the direst outcome: the demise of everyone. (Frankfurt 2018) Yet other contexts might suggest other normative goals, such as prioritarianism, which we will, however, not discuss in this paper.

These are situations where the particular context makes equality an unattractive normative goal that should be replaced by another pattern. Contextual pluralism seems to be necessary: According to a contextually pluralist approach, different patterns of justice may be most reasonable depending on features of the context, of the subjects affected, and the agents involved.

Criterion-Type and Optimization-Type Theories of Distributive Justice We propose a categorization of theories of distributive justice into *criterion-type* and *optimization-type* theories. Criterion-type theories of distributive justice have a well-defined normative goal, for which we can check whether it is fulfilled. These types of theories have counterfactual-invariant goals: We can always say whether either is achieved by observing the actual distribution and it is entirely irrelevant what other distributions could be realized. This is the case for egalitarianism and sufficientarianism. By contrast, maxmin requires us to compare the actual distribution to counterfactual ones (e.g., feasible alternatives) as its goal is the optimization of a value. How well we do can only be understood in comparison to possible alternatives. We will refer to this as optimization-type theories of distributive justice.[5] For the operationalization of theories of distributive justice, which we will get to later, it is important to understand these types.

[5]Other examples of this are prioritarianism and utilitarianism.

2.2 Group Fairness Criteria

Standard group fairness *criteria* such as statistical parity ($P(D = 1|A = a) = P(D = 1|A = b)$), equality of opportunity ($P(D = 1|Y = 1, A = a) = P(D = 1|Y = 1, A = b)$) (Hardt, Price, and Srebro 2016) or positive predictive value parity ($P(Y = 1|D = 1, A = a) = P(Y = 1|D = 1, A = b)$) demand equality in some metric across socio-demographic groups. What changes between them is *what* should be equal and for *whom* this should be equal. Statistical parity, for example, demands equal rates of positive decisions across groups while equality of opportunity also demands that but only for the parts of the group with a positive Y label. If we assume a *measure* M that is measured at the group level, these fairness criteria thus take the form $M(A = a) = M(A = b)$. In this paper, we take a utility-based approach (as previously suggested by Binns (2018); Finocchiaro et al. (2021); Jorgensen et al. (2023); Ben-Porat, Sandomirskiy, and Tennenholtz (2021); Hertweck, Loi, and Heitz (2024)), meaning that instead of comparing shares of positive decisions, we assume that we want to compare their consequences as measured by the utility U of a decision. We thus assume continuous (as opposed to strictly binary) utility values and compare these continuous values between groups. We take this approach for simplicity's sake to illustrate a more general point than it would have been possible with a binary utility.

2.3 Status of Conceptual Connection and Open Questions

Early work proposed group fairness criteria based on standard performance metrics such as accuracy and measures of the confusion matrix (see, e.g., Pedreshi, Ruggieri, and Turini (2008); Hardt, Price, and Srebro (2016); Verma and Rubin (2018); Corbett-Davies et al. (2023)). Starting in 2016, the infamous COMPAS case (Angwin et al. 2016) started a debate about the compatibility of fairness criteria (see Kleinberg, Mullainathan, and Raghavan (2016); Chouldechova (2017)). With that came the more urgent need for an understanding of fairness criteria from a philosophical and normative perspective. Attempts to connect group fairness criteria to philosophical theories mostly connected them to forms of egalitarianism. Gajane and Pechenizkiy (2017) and Heidari et al. (2019), for example, connected different group fairness criteria to different forms of egalitarianism while Binns (2018) gave a philosophical account of discrimination and egalitarianism to show "that 'fairness' as used in the fair machine learning community is best understood as a placeholder term for a variety of normative egalitarian considerations" (Binns 2018, p. 2). More recently, Kuppler et al. (2021) have criticized algorithmic fairness's focus on egalitarianism and discussed other theories of distributive justice that the literature could make use of. Some of these have already been operationalized as group fairness criteria (see, e.g., Martinez, Bertran, and Sapiro (2020); Diana et al. (2021)). However, in a lot of this work, the link between algorithmic fairness and distributive justice remains vague and underspecified. Mittelstadt, Wachter, and Russell (2023), for example, writes that "[group] fairness measures are related to egalitarian thinking in distributive justice" (Mittelstadt, Wachter, and Russell 2023, p. 19). It seems intuitive to connect distributive justice to algorithmic fairness as both are about the distribution of benefits/harms, but an important question remains open: *Theories of distributive justice typically operate at the individual level, but how do they relate to group fairness criteria that operate at the level of socio-demographic groups?* We claim that this unclarity is based on a fundamental misconception about how algorithmic fairness and distributive justice relate. We address this point with our rethinking of their relationship.

3 Rethinking Algorithmic Fairness and Distributive Justice

In this section, we present how we view distributive justice in relation to algorithmic fairness. We will also introduce the terminology that we will use in this context, such as distributive justice and structural injustice.

3.1 Differentiating Theories of Distributive Justice and Structural Injustices

Broadly speaking, our rethinking of distributive justice in the context of algorithmic decision-making builds on the hypothesis that decision-making systems are never perfect and that an ideal distribution cannot be reached in the real world, so we have to allow for deviations from perfect distributive justice. However, when we allow for deviations from the ideal, there is a chance that these individual deviations are unfairly distributed. It could be that a group is systematically more likely to be a victim of disadvantageous deviations. This is what we would then consider to be a *structural injustice* and what – so we argue – we want to test for in algorithmic fairness. As structural injustice is difficult to evaluate at the individual level, we use socio-demographic groups – this is the basis of group fairness criteria.

In theory, there are thus two options when we want to evaluate whether a theory of distributive justice is fulfilled:

1. Chosen theory of distributive justice is perfectly fulfilled (measured at the level of individuals) $\rightarrow$ no further tests necessary
2. Chosen theory of distributive justice is not perfectly fulfilled $\rightarrow$ (a) check if the theory of distributive justice is approximately fulfilled (measured at the level of individuals) and (b) check for signs of structural injustices (group fairness criteria measure this at the level of socio-demographic groups)

Thus, whenever our chosen theory of distributive justice cannot be perfectly fulfilled (which is essentially always the case in the real world),[6] we have to check for structural injustices.

[6]Note that egalitarianism, sufficientarianism and maximin are what Rawls (1999) calls "ideal theories" that assume ideal conditions, such as that every individual follows the rules. As Fazelpour and Lipton (2020) criticize, algorithmic fairness focuses on ideal theories of justice instead of non-ideal ones (which are needed in the real world).

In the context of our paper, we thus differentiate:

- **Theory of distributive justice:** Ideal distribution of benefits/harms among individuals
- **Structural injustices:** Unfair distribution of deviations from the perfect fulfillment of the theory of distributive justice among groups

Notice that for both the theory of distributive justice and structural injustices, we have to define *how* something should be distributed. The difference is in *what* is distributed: For theories of distributive justice, it is *benefits/harms themselves* that are distributed by a decision-making system. For structural injustices, it is the *deviations* from the ideal distribution of benefits/harms. In both cases, we have to define what the ideal distribution should look like – we could, e.g., ask for equality in benefits/harms or deviations. This is what we refer to as a **distributive pattern**. A distributive pattern does not say anything about what should be distributed or among whom it should be distributed – it is, as the name suggests, just a pattern for how something (let us call this X) should be distributed among individuals/groups/etc. (let us call these P). The distributive pattern can, e.g., be egalitarian (i.e., demanding equality of X among P), sufficientarian (i.e., demanding that the X of everyone in P reaches a certain threshold t) or maximin (i.e., demanding the maximization of the minimal X among P). Note that these are – confusingly – also the names of theories of distributive justice. The difference is that the respective theories of distributive justice do say something about what P is (typically individuals) and what X is.[7] Distributive patterns are thus used by both theories of distributive justice as well as structural injustices. Note that when we derive fairness criteria to test for structural injustices under different theories of distributive justice in Section 4, we will focus on the egalitarian pattern to measure structural injustices. However, Section 6 will discuss a possible alternative using the sufficientarian pattern.

3.2 Terminology: Criteria and Metrics

Before we dig in deeper, let us highlight the way in which we will use the terms *criterion* and *metric* in this paper, which are often used interchangeably in the literature. In our paper, we view *criteria* as binary conditions that can either be fulfilled or not. If a criterion is not perfectly fulfilled (which is to be expected), it is helpful to be able to quantify the deviation from this perfect state. This is what we will refer to as a *metric*. For example, if we expect equality in some *measure* M (i.e., we have an egalitarian criterion $\forall i, j \in P : M_i = M_j$), a possible metric is the absolute difference between the maximum and minimum values of the measures: $\max_{i \in P}(M_i) - \min_{j \in P}(M_j)$. Note that we can use a metric to create an *approximate criterion* by demanding that the criterion is sufficiently well fulfilled as measured by, e.g., the metric falling below a certain threshold t, e.g.,

[7]Although there are typically many variations of a theory of distributive justice. There are, e.g., different variations of egalitarianism that demand equality in resources, well-being or how we relate to people.

$\max_{i \in P}(M_i) - \min_{j \in P}(M_j) < t$. An issue is that when we derive a metric for a criterion (or the other way around), we always have multiple options for doing so. These translations emphasize different aspects of what the criterion (or metric) measures, which corresponds to subtle variations in our understanding of the criterion (or metric).

Note that we will use the terms criterion and metric both in the context of distributive justice, where we will call them *distributive justice criterion / metric*, and in the context of structural injustices, where we will call them *fairness criterion / metric*.[8] The following list gives an overview of this terminology:

- **Distributive justice criterion:** Measured at the individual level; tests whether the theory of distributive justice is perfectly fulfilled, which is presumably never the case in the real world
- **Distributive justice metric:** Measured at the individual level; measures deviation from perfect fulfillment of the theory of distributive justice
- **Fairness criterion:** Can be measured at different levels, but group fairness measures this at the level of socio-demographic groups; tests whether there are any signs of structural injustices, which we will probably always see in the real world as perfection is impossible to achieve.
- **Fairness metric:** Group fairness again measures this at the level of socio-demographic groups; measures deviation from perfect fulfillment of the fairness criterion

Note that for optimization-type theories (see Section 2.1), we cannot define a distributive justice criterion that does not take other feasible distributions into account – only a distributive justice metric. For criterion-type theories, the distributive justice criterion can be defined, but the metric is not unambiguous, and there are multiple options for how to define it. We will revisit this point when operationalizing maximin in Section 4.

3.3 Implications for Algorithmic Fairness

In current discussions of algorithmic fairness, the concepts of distributive justice, structural injustices and distributive patterns have frequently been mixed. We argue that algorithmic fairness has so far only looked at structural injustices but has not named this as such and has not made clear how this is related to theories of distributive justice.

When we want to test whether there are structural injustices in a given distribution – as we claim we do in algorithmic fairness – we thus have to define the following: (1) what theory of distributive justice the decision-making system is supposed to follow (because what structural injustice means depends on this contextual information) and then (2) how we define structural injustice in this context. Checking for perfect fulfillment of fairness criteria is not a realistic option in practice, so practitioners need both a distributive justice metric and a fairness metric. Existing group fairness criteria apply an egalitarian distributive pattern to test for structural injustices. However, note that we are not assuming that structural injustice is necessarily defined as the violation of an egalitarian pattern. We will deal with the question of how structural injustice could be measured with other patterns, e.g., in sufficientarian terms, in Section 6.

Fig. 1 outlines the relationship between testing for the fulfillment of a theory of distributive justice and potential structural injustices.[9]

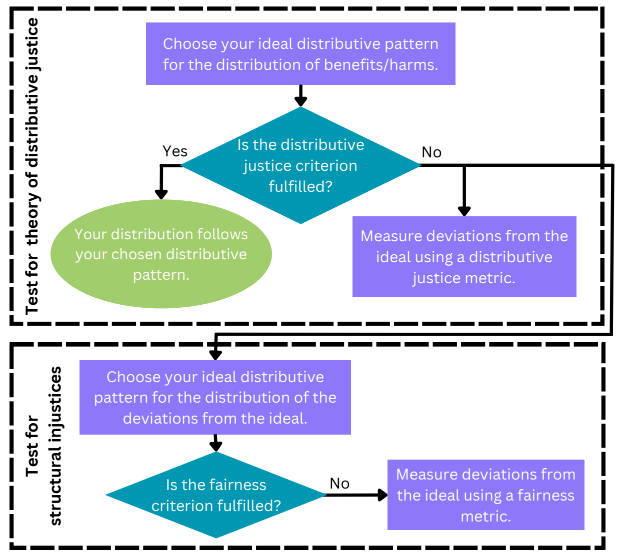


Figure 1: Partial flow chart to check the fulfillment of a theory of distributive justice and potential structural injustices in the deviations from that theory.

Ideally, we could demand both the chosen distributive justice metric and fairness metric to be optimized. However, they might be at odds and the question of how to balance these two is a moral one. When we have multiple possible models or distributions to choose from, it could make sense for practitioners to quantify the priority of the distributive justice metric compared to the fairness metric to help them choose a model if the better fulfillment of the theory of distributive justice and reduced structural injustice are at odds. One could, for example, imagine a Pareto front that shows the trade-offs between the two. One option would be to make the fairness criterion a constraint and then select the model that best fulfills the theory of distributive justice one has chosen.[10]

[8]We note that a more consistent term for our paper would be "structural justice criterion / metric", which we do not use in favor of the established terms "fairness criterion / metric".

[9]Note that only a part of the process is shown as there is no obvious procedure for this when a distributive pattern is not perfectly fulfilled. Ideally, both the distributive justice metric and the fairness metric show small enough deviations from the ideal, but these metrics might also be at odds – in which case trade-offs might be needed.

[10]This is what Rawls demands in his second principle (Rawls 1999): As Rawls puts fair equality of opportunity as a condition over the difference principle, his approach would be to filter for all distributions that do not show signs of structural injustices (viola-

Our takeaway from this is that whatever theory of distributive justice we believe our model should follow sets the basis for how we measure structural injustice: These two evaluations are not independent of each other. Rather, the measure of structural injustice depends on the chosen theory of distributive justice. The algorithmic fairness literature thus has to discuss fairness criteria / metrics in the context of their respective theory of distributive justice.

4 From Theories of Distributive Justice to Fairness Criteria

In this section, we will show how the way in which we view distributive justice impacts algorithmic fairness. Specifically, we will go over three theories of distributive justice (egalitarianism, sufficientarianism, and maximin) and discuss how to check the fulfillment of the theory of distributive justice as well as potential structural injustices.

4.1 Egalitarianism

We start with egalitarianism as this is the theory of distributive justice that group fairness criteria have retroactively been tied to (see Section 2).

Test for Theory of Distributive Justice

Distributive Justice Criterion If egalitarianism is perfectly fulfilled, everyone gets exactly the same. We can, therefore, evaluate egalitarianism as a theory of distributive justice at the individual level with the following justice criterion:

Proposition 1. The **distributive justice criterion for egalitarianism** is that all (equally deserving) individuals have the same utility: $\forall i, j \in P : u_i = u_j$.

Distributive Justice Metric If the justice criterion is not perfectly fulfilled, our rethinking of distributive justice has shown that we can check whether the distributive pattern is approximately fulfilled. When we want to check whether our distribution fulfills egalitarianism well *enough*, we could come up with a metric that measures the deviation from perfect fulfillment. Examples of possible **distributive justice metrics** for egalitarianism are:

- The Gini coefficient (Gini 1912) is a popular metric to measure inequality that is usually used to measure wealth or income inequality, e.g., in a country
- Variance in utilities across the entire population: $\mathrm{Var}(U)$
- Difference between the maximum and minimum utility: $\max_{i \in P} u_i - \min_{j \in P} u_j$
- Ratio of the maximum and minimum utility: $\dfrac{\max_{i \in P} u_i}{\min_{j \in P} u_j}$

tions of the fair equality of opportunity principle) and to then select the one that best fulfills the difference principle. He calls this the lexical priority of fair equality of opportunity relative to the difference principle: "The second principle of justice is lexically prior to the principle of efficiency and to that of maximizing the sum of advantages; and fair opportunity is prior to the difference principle" (Rawls 1999, p. 266).

We could then set a threshold for the metric to define how far the distribution can deviate from perfection to still be called "approximately just". This gives us an **approximate distributive justice criterion**. The aim is to say "While not every utility is exactly equal, the utilities are similar enough according to our chosen metric."

Test for Structural Injustices

As we know, when the theory of distributive justice is not perfectly fulfilled, this gives rise to potential biases. This is what we would then consider to be a sign of structural injustice, which we have to check for separately.

Let us use two examples to figure out what structural injustices would look like in this context. Example 1 shows what structural injustice could look like for a distribution that approximately fulfills the theory of distributive justice. Example 2 then highlights that a distribution that does not exhibit these signs of structural injustice does not have to be egalitarian though – thereby showing that there is indeed a difference in checking for an approximate just distribution according to egalitarianism and checking for the absence of structural injustice.

- **Example 1:** Assume that individuals are graded on an exam. We pick the students that are equally deserving of a B to check the grading process. With this background information, everyone in our set should get a B. Assume further that there are two groups, group 1 and group 2. The teacher is biased against group 2, which means they give every group 2 student a C but every group 1 student the B they deserve. Of course, the individual-level check for egalitarianism would fail as not every student gets the same grade. However, if we assume group 2 is small in size, the Cs might just be a few outliers that do not influence the variance much. We might thus say that the distribution is just. However, we cannot tell yet if there are structural injustices. For this, we have to check the grades of group 1 and group 2. We could, for example, compare grade averages to realize that the teacher systematically grades group 2 students lower than group 1 students. We thus need the fairness criterion to check for structural injustice.
- **Example 2:** Assume that instead of distributing bank loans based on an applicant's properties, a bank clerk just tosses a coin. Everyone has a 50-50 chance of getting the loan. While there is certainly something to be said about procedural justice here (a loan grant should probably not be decided based on a coin toss), let us focus on outcome fairness for now. Here, it seems that the resulting loans are not fair on an individual level if we think that getting a loan should at least be somewhat related to whether we can repay the loan or some other individual characteristics. If we focused on equally deserving individuals (however deservingness is defined), we see that there is high variance, so our distribution would not be seen as just by the distributive justice metric. However, we also know for sure that whether you get a loan or not is not causally influenced by group membership – there is thus no structural injustice, which would also be reflected by a fairness criterion comparing group averages.

These two examples show that testing for the theory of distributive justice and testing for structural injustices are two different things and that both are necessary.

Fairness Criterion Fairness criteria are supposed to detect possible structural injustices, which we defined as unfairly distributed deviations from an ideal distribution. We are thus looking for patterns of systematic differences between groups by comparing the expected utilities of both groups. A (notable) difference in these expectation values hints at possible structural injustices.

Proposition 2. A possible **fairness criterion for egalitarianism** is that the expected utilities of all groups are equal: $\forall a, b \in A : E(U|a) = E(U|b)$.

This is essentially how standard group fairness criteria operate.[11] This tells us that standard group fairness criteria seem to assume egalitarianism as the theory of distributive justice and then also use the egalitarian distributive pattern to measure potential structural injustices.

Fairness Metric Since this fairness *criterion* is almost impossible to perfectly fulfill in practice, we also want to measure the deviation from the criterion, which we formalize as *metrics*. The list of possible metrics is similar to the list of possible distributive justice metrics: the Gini coefficient, variance, a difference or ratio, etc. This similarity occurs because both criteria (the distributive justice criterion and the fairness criterion) demand equality in something (utilities/-expected utility) across a set (of individuals/groups).

4.2 Sufficientarianism

We now apply our approach to sufficientarianism to derive appropriate fairness criteria and metrics from this. This will give us to an important insight: To check for structural injustices in a sufficientarian distribution, we still use an egalitarian pattern – what changes compared to the fairness criteria for egalitarian distributions is the measure that we demand to be equal.

Test for Theory of Distributive Justice

Distributive Justice Criterion If sufficientarianism is perfectly fulfilled, everyone's utility is above the predefined minimum threshold:

Proposition 3. The **distributive justice criterion for sufficientarianism** is that all (equally deserving) individuals have a utility above the threshold t: $\forall i \in P : u_i > t$.

Distributive Justice Metric An intuitive distributive justice metric is the share of individuals with a utility above the threshold t. To get an approximate distributive justice criterion, this metric can be demanded to be sufficiently high.

Test for Structural Injustices

We again discuss two examples to show that an approximately just distribution does not have to go hand in hand with the absence of structural injustices to help us understand what would constitute structural injustices in the context of sufficientarianism.

[11]Except that they do not take a utility-based approach and instead compare binary decisions or outcomes across groups.

- **Example 1:** Picture a world in which 95% of individuals reach the threshold, which is assumed to be sufficient for the distribution to be seen as approximately fulfilling sufficientarianism. However, the 5% not reaching the threshold are all in group 1 while the 95% reaching the threshold are in group 2. There is clearly an issue of structural injustice despite the approximate distributive justice test passing.
- **Example 2:** Imagine that approximately equal shares of group 1 and 2 are above the threshold, so that there is no issue of structural injustice. Now imagine, however, that overall only 1% of the entire population is above said threshold. Sufficientarianism is then not reached even though there is no issue of structural injustice.

Fairness Criterion Sufficientarianism assumes that there are utility thresholds that, ideally, every individual should reach. This could, for example, be the case in the distribution of medicine, where everyone needs at least one pill to be healed (half of a pill does not work). Then the threshold is one pill. It does not matter whether an individual receives more than one pill – they will be healed either way. Likewise, it does not matter whether they receive half a pill or no pill as they will stay sick. In such a setting, what are deviations from the ideal distribution? Under the ideal distribution, everyone reaches the threshold. Deviations are cases where people do not reach the threshold. It does not matter how far they fall below the threshold, so for the deviations, we only care whether they are below the threshold; this is thus a binary measurement. Then, the question of structural injustice is: What is an unfair distribution of these deviations? We propose the following fairness criterion:

Proposition 4. A possible **fairness criterion for sufficientarianism** is that the share of individuals with utilities above the threshold t_u is equal across all groups: $\forall a, b \in A : E(U'|a) = E(U'|b)$, where $U' = I(U > t)$, so U' models the resulting utility as binary: either one is above the threshold ($U' = 1$) or below ($U' = 0$).

Here, we notice that checking for structural injustices under sufficientarianism leads us to an egalitarian distributive pattern: To check if there is a systematic bias against a group, we check if groups are (approximately) treated equally. What changes compared to Section 4.1 is *what* we demand to be equal: It is not $E(U)$ but $E(U')$. One can then define *fairness metrics* and *approximate fairness criteria* in the same manner as we did for the egalitarian theory of distributive justice in Section 4.1.

4.3 Maximin

Let us now apply this approach to maximin. Remember that maximin is an optimization-type theory of distributive justice (see Section 2.1), which has an influence on the operationalization.

Test for Theory of Distributive Justice

Distributive Justice Metric If we follow maximin, the utility of the worst-off individual $min_{i \in P}(u_i)$ is maximized. However, assuming that decision-making systems are not perfect, we might not want to focus on individuals. To avoid looking at the single worst-off individual (there could always be an outlier), we propose measuring the expected utility of the worst-off 5% of individuals. Note that here, we only derive a metric and not a criterion due to maximin being an optimization-type theory of distributive justice.[12]

Test for Structural Injustices
How does structural injustice play out for maximin? Imagine the following two distributions of wages per hour of work where all people whose names start with the letter "A" are in group A and all people whose names start with the letter "B" are in the marginalized group B. Assume that group A and B are roughly equal in their natural talents and motivation relevant to the job and that the distributions exhibited here are statistically significant, so we assume that the inequality is due to unequal opportunities.

- **A**nna: 10€, **B**erta: 20€, **A**nton: 30€, **B**asti: 40€, **A**driana: 50€, **B**arbara: 60€
- **B**erta: 10€, **B**asti: 20€, **B**arbara: 30€, **A**nna: 40€, **A**nton: 50€, **A**driana: 60€

Both distributions are equally good according to maximin as the worst-off person always has 10€ (and using the approximate version, the worst-off 5% are also equal in their utility). Given that maximin cares about the worst-off individual, this theory of justice assigns moral value to the one that is worst-off. This gives us reason to believe that someone who subscribes to maximin and cares about structural injustices should find the first distribution preferable to the second one.[13] There are multiple possible explanations as to why they should prefer the first distribution:

- The worst-off individuals of the second distribution are all in group B.
- The worst-off individuals of group B (Berta, Basti) are worse off than the worst-off individuals of group A (Anna, Anton) in the second distribution.

To evaluate structural injustices, we do not just check what group the single worst-off individual belongs to: of course, this individual has to belong to some group, and this does not yet constitute a structural injustice against that group. We argue that it only becomes problematic once we see a pattern at the group level. We, therefore, look at the 5% worst-off individuals (in the population/per group) – similar to the distributive justice metric.

Fairness Criterion From this, we could come up with different fairness criteria:

- Fairness demands that it is (approximately) equally likely for members of each group to be in the worst-off 5%.
- Fairness demands that the expected utility of the worst-off 5% in each group have an (approximately) equal expected utility.

Again, these fairness criteria only tell us something about fairness and not about which distribution is more just according to maximin. Imagine, for example, a distribution where half of the worst-off 5% are in group A and the other half in group B – or a distribution where the worst-off 5% of group A have the same expected utility as the worst-off 5% of group B. A second distribution does not fulfill this, but the worst-off individual is (or the worst-off 5% are) better off in distribution 2. In that case, our fairness criterion would tell us to prefer distribution 1, but the distributive justice metric would actually require distribution 2. This again shows that the criteria and metrics that we use to evaluate the justness of a distribution are different from the criteria and metrics that we use to evaluate structural injustices.

Analogous to fairness criteria for an egalitarian (Section 4.1) and a sufficientarian (Section 4.2) theory of distributive justice, we can turn the fairness criteria for maximin into **fairness metrics** and **approximate fairness criteria**. Note that by picking two options to translate maximin to a fairness criterion, we get a multitude of fairness metrics as there are also multiple options for translating each of these fairness criteria into a metric. What does this tell us and how should we choose between them? We argue that they highlight different aspects of what matters about maximin and structural injustices. Depending on one's views on this, one can choose different paths and end up with a different criterion or metric.

5 Comparison to Naive Operationalizations

As previously discussed, examples to illustrate theories of distributive justice are often given at the individual level. However, group fairness criteria are defined at the group level. Our approach shows fairness criteria depend on the theory of distributive justice they were created for. This section will highlight the necessity of our approach by comparing the resulting fairness criteria to a more naive approach that simply takes the theory of justice defined at the individual level and replaces the term "individual" with "group" and "utility" with "expected utility". Such an approach would result in the following fairness criteria:

- **Egalitarianism:** Every individual has the same utility. → Every group has the same expected utility.
- **Sufficientarianism:** Every individual's utility is above the threshold. → Every group's expected utility is above the threshold.
- **Maximin:** Maximize the utility of the worst-off individual. → Maximize the expected utility of the worst-off group.

These naive operationalizations of theories of justice as group-level fairness criteria are a tempting way to test for structural injustice – after all, Section 4.1 showed that it seems to work for egalitarianism. In this section, we will discuss why the naive operationalization of theories of justice as group-level fairness criteria can turn out to be logi-

[12]We could turn it into a criterion by asking that among all possible distributions R, we choose the one that maximizes our metric.

[13]Note that this cannot be avoided by demanding a leximin distribution as the two distributions are also equally good according to leximin. Yet, since leximin cares about those who are worst off, someone who cares about structural injustices should again prefer the first distribution.

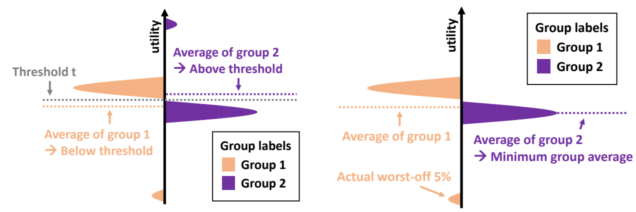
cally inconsistent with said theory of distributive justice and mismatch our intuitions about said theory.

5.1 Sufficientarianism

Our approach to operationalizing sufficientarianism results in a different fairness criterion than the naive operationalization.[14] As we will show in this section, the naive operationalization is not logically coherent with the idea of sufficientarianism. To see this, imagine two groups. In group 1, 95% of the members are above the threshold. In group 2, only 5% are above the threshold. This is illustrated in Fig. 2a. Now imagine further that the 95% above the threshold in group 1 are just barely above the threshold while the other 5% are far below the threshold. The resulting average group utility is below the threshold, so the naive operationalization of sufficientarianism would constitute that group 1 does not fulfill the sufficientarian requirement. In group 2, however, the 95% below the threshold are barely below the threshold while the other 5% are far above the threshold. Here, the resulting average utility is above the threshold, meaning the naive operationalization would see sufficientarianism to be fulfilled – despite 95% of the group not reaching the threshold. Assuming that the threshold represents what is needed to survive, this naive operationalization would assume that group 2 is better off than group 1 even though 95% of individuals in group 2 are starving while it is "only" 5% in group 1. This means that this naive operationalization does not tell us anything about structural injustices. Clearly, group 1 is better off according to sufficientarianism and group 2 is the disadvantaged group. Our fairness criterion derived in Section 4.2, which checks if approximately equal shares from all groups are above the threshold, accounts for this. Moreover, this naive operationalization is not even fit to check if sufficientarianism is fulfilled. This is where our distributive justice metric is necessary, which checks if a large enough share of the population reaches the threshold.

5.2 Maximin

Both Heidari et al. (2019)'s[15] and Martinez, Bertran, and Sapiro (2020)'s[16] maximin fairness metric correspond to the naive version of translating maximin to the socio-demographic group setting that we described at the beginning of Section 4. The issue with this operationalization is that if the worst-off individuals are in a group with a lot of very well-off individuals, then the actual worst-off individuals are ignored in the distribution process. Imagine, for example, that the 5% worst-off individuals are distributed across groups that are, on average, fairly well-off as seen for group 1 in Fig. 2. Now imagine that another group (group 2 in Fig. 2) has a lower average utility, but the worst-off individuals from that group are doing much better than the worst-off individuals from group 1. The worst-off individuals from group 1 are then ignored as we only maximize the utility of the group that is, on average, worst-off, so group 2. As the previous discussion of how to evaluate structural injustice for maxmin showed, there are other fairness criteria that are – so we argue – more in the spirit of maximin.



(a) Sufficientarianism (b) Maximin

Figure 2: Illustration of the issues that can arise through the naive operationalization

6 Is Group Fairness Inherently Egalitarian?

In Section 4, we defined fairness criteria under different theories of distributive justice. This helps us check whether a decision-making system treats people of a disadvantaged group unfairly. Notice that for the sufficientarian theory of distributive justice, we demanded *equality* in the share of groups' members above the threshold t while for maximin, we proposed two fairness criteria that both demand *equality* in a measure between groups. One might ask if there are alternatives to this egalitarian fairness approach – in particular because this type of egalitarianism still carries the risk of leveling down: When such a fairness criterion is enforced, it could actually happen that the share of members above the threshold is lowered in all groups (for sufficientarianism) or that, e.g., the expected utility of the worst-off individuals is lowered (for maximin). Mittelstadt, Wachter, and Russell (2023) have suggested a fairness notion, which does not check for equality but instead demands some measure to reach a certain level. They call this type of fairness constraint "leveling up" and describe it as follows: "[If] we believe that people are being harmed by low selection rates, precision, or recall, instead of enforcing that these properties be equalised across groups, we can instead require that every group has, at least, a minimal selection rate, precision, or recall." (Mittelstadt, Wachter, and Russell 2023, p. 37)

[14] One might be tempted to think that Mittelstadt, Wachter, and Russell (2023) suggests such a naive operationalization of sufficientarianism: instead of demanding that all individuals reach a certain threshold, Mittelstadt, Wachter, and Russell (2023)'s fairness criterion demands that all groups reach a certain threshold. However, we argue that Mittelstadt, Wachter, and Russell (2023) should actually be understood differently. Section 6 will therefore discuss a more charitable interpretation of Mittelstadt, Wachter, and Russell (2023)'s work as well as how their work fits into our understanding of algorithmic fairness.

[15] Heidari et al. (2019) briefly discusses maximin as an alternative to egalitarianism citing the leveling down objection. They suggest that "the max-min distribution deviates from equality only when this makes the worst off group better off", but do not discuss why maximin can or should be analyzed at the group level.

[16] Martinez, Bertran, and Sapiro (2020) integrates maximin into a formalized fairness criterion, which has already been used in other influential papers such as Diana et al. (2021). However, the paper does not give a moral justification for their operationalization.

As mentioned in Section 5.1, this is reminiscent of sufficientarianism. However, Mittelstadt, Wachter, and Russell (2023) never portray their proposed fairness criterion as an operationalization of sufficientarianism. Instead, they portray it in a way that is analogous to the way egalitarianism is used in group fairness criteria: Not as a criterion to check if this theory of distributive justice is fulfilled but rather as a fairness criterion to check if there are structural injustices at the group level. In fact, "leveling up" does not represent a single fairness criterion but contains multiple possible fairness criteria: Leveling up does not specify *what* ought to be above the minimum threshold[17] – just how egalitarianism does not specify *what* ought to be equal. Therefore, leveling up could ask for, e.g., the selection rates to be above a threshold t_{LU} ($\forall a \in A : P(D = 1|A = a) > t_{LU}$) or for the true positive rate to be above the threshold ($\forall a \in A : P(D = 1|Y = 1, A = a) > t_{LU}$) – just like egalitarianism contains multiple fairness criteria depending on what measure M is chosen. A formalization of this making use of the generic measure M and the threshold t_{LU} is $\forall a \in A : M(A = a) > t_{LU}$.

We can combine our operationalization of fairness criteria for egalitarianism, sufficientarianism and maximin with Mittelstadt, Wachter, and Russell (2023)'s leveling up. In the case of the egalitarian theory of distributive justice, we can demand that the expected utilities of groups reach a minimum threshold t_{LU}: $\forall a \in A : E(U|A = a) > t_{LU}$. Instead of demanding equal expected utilities, we would demand a minimum level for the expected utility of each group. In the case of sufficientarianism, we can demand the share of individuals with utilities above the threshold t to reach a certain threshold t_{LU} for all groups: $\forall a \in A : E(U > t|A = a) > t_{LU}$ instead of demanding the share of individuals above the threshold t to be equal across groups ($E(U > t|A = a) = E(U > t|A = b)$). This means that while we want equal shares of members from all groups to reach the threshold, we can now forgo equality if it harms a group and instead just demand that groups bring at least a certain share t_{LU} of individuals above the required threshold t. This notion is useful if you believe that it is, e.g., sufficient to get 80% of individuals above the threshold in every group and it does not matter if one group gets significantly more of their members above the threshold than another group. If you, however, believe that this is a sign of structural injustice, you should use the egalitarian distributive pattern to test for structural injustices. Similarly, we can adapt our fairness criteria for maximin. For the second fairness criterion (the expected utility of the worst-off 5% in each group should be (approximately) equal across groups), we could, for example, demand that the expected utility of the worst-off 5% in each group reaches a minimum threshold t_{LU}.

There is thus a question of what we mean by structural injustice: Do we speak of structural injustice when the harms of deviating from a distribution that fulfills our chosen theory of distributive justice are unequally distributed? This would require applying the egalitarian pattern to structural injustice. Or do we, rather, want to ensure that harmful deviations remain small for each group? Then, we might want to take a sufficientarian approach to structural injustice (i.e., Mittelstadt, Wachter, and Russell (2023)'s "leveling up"). When choosing a group-level fairness criterion or metric, we thus first have to decide on what theory of distributive justice we find most appropriate. Then, we need to decide what we view as structural injustices under the chosen type of distribution.[18] However, one could reasonably disagree on this question as it does not have a straightforward answer and depends on how one views what constitutes structural injustices.

[17]They mention multiple options such as the true positive rate, the false positive rate, etc. While their examples focus on comparing binary decisions across groups, they do not state their approach as being limited to binary decisions. We, therefore, interpret their proposal as being extendable to compare utility values.

7 Conclusion

Group fairness criteria are typically described as "egalitarian". A lot of the algorithmic fairness literature, therefore, seems to assume that fairness criteria are the operationalization of theories of distributive justice. However, we argue that this is not true: While theories of distributive justice demand a certain distribution of benefits/harms among individuals, we claim that fairness criteria actually test whether the inevitable deviations from said ideal are unfair to certain groups, i.e., systematically biased against them. This test for a systematic bias in the deviations often uses an egalitarian distributive pattern and is therefore confused with the egalitarian theory of distributive justice. We show how this refined conceptual thinking affects what fairness criteria we use when we have a distribution that is supposed to follow theories of distributive justice other than egalitarianism.

Implications In current discussions of algorithmic fairness, the concepts of theories of distributive justice, structural injustices and distributive patterns are often confused. Yet, recognizing these concepts and their relationships is crucial as the measurement of structural injustices depends on the chosen theory of distributive justice. When testing for structural injustices, this means first defining the theory of distributive justice and also testing whether it is fulfilled. When proposing new fairness criteria, this means discussing the theory of distributive justice in the context in which these fairness criteria can be used. Our paper is thus a call for the algorithmic fairness literature to differentiate between theories of distributive justice and structural injustices.

Future work Future work should look into the subtle normative differences between the multiple fairness metrics that can be derived from a single fairness criterion and the other way around. This could provide valuable guidance to practitioners. It would also be interesting to apply the demonstrated approach to derive fairness criteria for other theories of distributive justice such as prioritarianism.

[18]Rawls (1999)'s second principle, for example, tells us that he views group-level fairness for maximin as egalitarian – represented by what he calls "fair equality of opportunity".

Acknowledgements

Thank you to Joachim Baumann for his valuable feedback and input on several versions of this draft. We also want to thank Marcello Di Bello for his thoughtful feedback on this paper. This work was supported by the National Research Programme "Digital Transformation" (NRP 77) of the Swiss National Science Foundation (SNSF), grant number 187473.

References

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine Bias. *ProPublica*.

Ben-Porat, O.; Sandomirskiy, F.; and Tennenholtz, M. 2021. Protecting the protected group: Circumventing harmful fairness. In *Proceedings of the aaai conference on artificial intelligence*, volume 35, 5176–5184.

Binns, R. 2018. Fairness in machine learning: Lessons from political philosophy. In *Conference on fairness, accountability and transparency*, 149–159. PMLR.

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

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

Cohen, G. A. 1990. Equality of what? On welfare, goods and capabilities. *Recherches Économiques de Louvain/Louvain Economic Review*, 56(3-4): 357–382.

Corbett-Davies, S.; Gaebler, J. D.; Nilforoshan, H.; Shroff, R.; and Goel, S. 2023. The measure and mismeasure of fairness. *The Journal of Machine Learning Research*, 24(1): 14730–14846.

Diana, E.; Gill, W.; Kearns, M.; Kenthapadi, K.; and Roth, A. 2021. Minimax group fairness: Algorithms and experiments. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 66–76.

Fazelpour, S.; and Lipton, Z. C. 2020. Algorithmic fairness from a non-ideal perspective. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 57–63.

Finocchiaro, J.; Maio, R.; Monachou, F.; Patro, G. K.; Raghavan, M.; Stoica, A.-A.; and Tsirtsis, S. 2021. Bridging machine learning and mechanism design towards algorithmic fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 489–503.

Frankfurt, H. 2018. Equality as a moral ideal. In *The Notion of Equality*, 367–389. Routledge.

Gajane, P.; and Pechenizkiy, M. 2017. On formalizing fairness in prediction with machine learning. *arXiv preprint arXiv:1710.03184*.

Gini, C. 1912. *Variabilità e mutabilità: contributo allo studio delle distribuzioni e delle relazioni statistiche.[Fasc. I.]*. Tipogr. di P. Cuppini.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

Heidari, H.; Loi, M.; Gummadi, K. P.; and Krause, A. 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.

Hertweck, C.; Loi, M.; and Heitz, C. 2024. Group Fairness Refocused: Assessing the Social Impact of ML Systems. In *2024 11th Swiss Conference on Data Science (SDS)*, 189–196. IEEE.

Jacobs, A. Z.; and Wallach, H. 2021. Measurement and fairness. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 375–385.

Jorgensen, M.; Richert, H.; Black, E.; Criado, N.; and Such, J. 2023. Not so fair: The impact of presumably fair machine learning models. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 297–311.

Kagan, S. 2012. The Ratio View. In *The Geometry of Desert*. Oxford University Press.

Kasirzadeh, A. 2022. Algorithmic fairness and structural injustice: Insights from feminist political philosophy. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 349–356.

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

Kuppler, M.; Kern, C.; Bach, R. L.; and Kreuter, F. 2021. Distributive justice and fairness metrics in automated decision-making: How much overlap is there? *arXiv preprint arXiv:2105.01441*.

Lamont, J.; and Favor, C. 2017. Distributive Justice. In Zalta, E. N., ed., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2017 edition.

Loi, M.; Herlitz, A.; and Heidari, H. 2023. Fair equality of chances for prediction-based decisions. *Economics & Philosophy*, 1–24.

Martinez, N.; Bertran, M.; and Sapiro, G. 2020. Minimax pareto fairness: A multi objective perspective. In *International Conference on Machine Learning*, 6755–6764. PMLR.

Mittelstadt, B.; Wachter, S.; and Russell, C. 2023. The Unfairness of Fair Machine Learning: Levelling down and strict egalitarianism by default. *arXiv preprint arXiv:2302.02404*.

Pedreshi, D.; Ruggieri, S.; and Turini, F. 2008. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 560–568.

Rawls, J. 1999. *A Theory of Justice*. Cambridge, Massachusetts: Harvard University Press, 2 edition.

Sen, A. 1980. Equality of what? *The Tanner lecture on human values*, 1: 197–220.

Temkin, L. 2017. Equality as comparative fairness. *Journal of Applied Philosophy*, 34(1): 43–60.

Verma, S.; and Rubin, J. 2018. Fairness definitions explained. In *Proceedings of the international workshop on software fairness*, 1–7.

Young, I. M. 2010. *Responsibility for justice*. Oxford University Press.