Assessing Group Fairness with Social Welfare Optimization

Violet Chen[1], J. N. Hooker[2(✉)], and Derek Leben[2]

[1] Stevens Institute of Technology, Hoboken, USA
vchen3@stevens.edu
[2] Carnegie Mellon University, Pittsburgh, USA
{jh38,dleben}@andrew.cmu.edu

Abstract. Statistical parity metrics have been widely studied and endorsed in the AI community as a means of achieving fairness, but they suffer from at least two weaknesses. They disregard the actual welfare consequences of decisions and may therefore fail to achieve the kind of fairness that is desired for disadvantaged groups. In addition, they are often incompatible with each other, and there is no convincing justification for selecting one rather than another. This paper explores whether a broader conception of social justice, based on optimizing a social welfare function (SWF), can be useful for assessing various definitions of parity. We focus on the well-known alpha fairness SWF, which has been defended by axiomatic and bargaining arguments over a period of 70 years. We analyze the optimal solution and show that it can justify demographic parity or equalized odds under certain conditions, but frequently requires a departure from these types of parity. In addition, we find that predictive rate parity is of limited usefulness. These results suggest that optimization theory can shed light on the intensely discussed question of how to achieve group fairness in AI.

Keywords: Social welfare optimization · group parity in AI

1 Introduction

There is growing demand within industry and government for assurance that machine learning (ML) models respect and promote equality of impact across protected groups [18,25] and comply with legal requirements [17,42]. This concern arises in contexts that range from hiring and parole decisions to mortgage lending and credit ratings. One prominent method of satisfying these ethical and legal goals is the use of statistical parity metrics [3]. For example, one might assess two groups have equal approval rates (*demographic parity*), whether the approval and rejection rates of qualified candidates are equal (*equalized odds*), or whether the fraction of qualified candidates among those approved is the same (*predictive rate parity*).

There are at least two problems, however, with reliance on statistical parity as a measure of fairness. One is that parity metrics take no account of the actual

B. Dilkina (Ed.): CPAIOR 2024, LNCS 14742, pp. 208–225, 2024.
https://doi.org/10.1007/978-3-031-60597-0_14

utility consequences of being selected or rejected. Presumably, group disparities are viewed as unjust because different groups derive unequal benefits from the selection process. Yet an assessment of these benefits requires consideration of the actual welfare outcomes of selecting or rejecting individuals. For example, rejecting a member of a disadvantaged group may have greater negative consequences than rejecting a member of an advantaged group. The standard parity metrics take account only of the number of individuals selected or rejected, not the impacts of these decisions.

A second problem is that parity metrics are frequently incompatible with each other [14,19,29] and, in particular, imply different trade-offs between fairness and accuracy [5,29]. As a result, there is often no consensus on which metric is appropriate in a given context. This is illustrated by the famous debate over parole decisions between ProPublica and Northpointe (now Equivant) regarding whether the latter's COMPAS product is fair, with one side claiming that the model is unfair because it fails to achieve equalize odds, and the other side claiming it is fair because it achieves predictive rate parity [1,16]. Lacking any further grounds for settling this dispute, the debate has (for now) reached a stalemate. Ideally, one would justify (or reject) a parity metric by appealing to a broader principle of justice.

In this paper, we explore an approach for evaluating group parity metrics via their effects on the welfare of individuals in each group. The aim is to connect the debate about group parity with the rich tradition of welfare economics, where policies are evaluated by their effects on social welfare, as measured by a social welfare function (SWF). Such a function can take into account the distribution of utilities as well as overall welfare. We ask whether a selection policy that optimizes social welfare, as measured by a SWF, results in some particular form of group parity or requires departure from the standard parity measures. Our underlying hypothesis is that insights obtained from optimization theory can shed light on the vexing problem of fairness in AI.

As a first step in this research program, we investigate the parity implications of *alpha fairness* [36,43], a well-known family of SWFs parameterized by a nonnegative real number α. Larger values of α indicate a stronger emphasis on fairness as opposed to maximizing total utility, the latter corresponding to $\alpha = 0$. Alpha fairness can therefore evaluate the trade-off of fairness and accuracy, a perennial issue in machine learning. Other special cases include the maximin (Rawlsian) criterion ($\alpha = \infty$) and *proportional fairness*, also known as the Nash bargaining solution ($\alpha = 1$). We ask what are the parity implications of a given level of fairness as indicated by α.

Our purpose here is not to defend alpha fairness as a fairness criterion, but to explore the implications of a criterion that has *already* been extensively defended. Alpha fairness in its various forms has been studied for over 70 years by investigators that include two Nobel laureates (John Nash and J. C. Harsanyi). Nash [38] gave an axiomatic argument for his bargaining solution in 1950, while Rubinstein, Harsanyi and Binmore [6,21,41] supplied bargaining arguments. Lan et. al [30,31] provided an axiomatic derivation for general alpha fairness and proposed an interpretation of the α parameter. Bertsimas et al. [5] studied resulting equity/efficiency trade-offs. Alpha fairness has also seen a number of practical applications, particularly in telecommunications and other engineering fields [28,34,36,39,43].

After a brief survey of related work, we first establish a general solution to the problem of maximizing alpha fairness subject to a constraint on the number of individuals selected. We then present a utility model that allows us to relate group characteristics to the implications of alpha fairness. Following this, we describe specific implications for demographic parity, equalized odds, and predictive rate parity, and draw conclusions from these results.

2 Related Work

Statistical group parity metrics are the most widely studied approach to fairness in AI and machine learning. Much of this research is surveyed in [10,35]. However, a welfarist approach is beginning to receive recognition in AI fairness literature, e.g. [4,8,11,13,15,23,24,33]. One motivation is pragmatic: social welfare can provide a "common currency" with which one can justify the choice of parity metric, when the typical justifications are incommensurate [20]. For example, arguments for individual fairness appeal to procedural justice concerns, while arguments for group fairness appeal to distributive justice [7,32]. When a model cannot satisfy both of these values, it is necessary to justify one's choice. Another motivation for a welfarist approach is ethical: one may wish to strive for group parity to make disadvantaged groups better off, rather than to achieve equality for its own sake [9,37].

Social welfare functions have been used in optimization models for some time, as surveyed in [13,27,40]. Aside from alpha fairness, SWFs that balance equity and efficiency include Kalai-Smorodinsky bargaining [26] and threshold functions [12,22,44].

Despite the large literature on SWFs and group parity metrics, we describe here what is, to our knowledge, the first explicit connection between them.

3 The Basic Model

We address the task of selecting individuals from a population to receive a benefit or resource, such as a mortgage loan or a job interview. Some individuals belong to a protected group that is disadvantaged with respect to qualification status. We define binary variables D, Y, Z to indicate whether an individual is selected ($D = 1$), qualified ($Y = 1$), or protected ($Z = 1$). To simplify notation, we use D to represent $D = 1$ and $\neg D$ to represent $D = 0$, and similarly for Y and Z.

We have demographic parity when $P(D|Z) = P(D|\neg Z)$, equalized odds (in the positive sense of equality of opportunity) when $P(D|Y,Z) = P(D|Y,\neg Z)$, and predictive rate parity when $P(Y|D,Z) = P(Y|D,\neg Z)$. We interpret the conditional probability $P(D|Z)$ as the fraction of protected individuals who are selected, and similarly for the other probabilities. The latter two types of parity are typically defined in terms of qualifications that are determined after the fact, such as whether a mortgage recipient repaid the loan, a job interviewee was hired, or a parolee committed no further crimes. In addition, calculation of the odds ratio requires knowledge of how many rejected candidates are qualified.

To assess utilitarian outcomes, we suppose that an individual i experiences expected utility $u_i = a_i + b_i$ if selected, and a baseline utility $u_i = b_i$ if rejected. We refer to a_i as the *selection benefit*. It can be negative (indicating that selection is harmful), but we assume that $b_i > 0$ and $a_i + b_i > 0$ because alpha fairness is not defined for nonpositive utilities. This assumption can be met by a positive translation of the utility scale if necesssary.

We assess the desirability of a utility distribution $\boldsymbol{u} = (u_1, \ldots, u_n)$ with the alpha fairness social welfare function, given by

$$W_\alpha(\boldsymbol{u}) = \begin{cases} \dfrac{1}{1-\alpha} \displaystyle\sum_i u_i^{1-\alpha}, & \text{if } \alpha \geq 0 \text{ and } \alpha \neq 1 \\ \displaystyle\sum_i \log(u_i), & \text{if } \alpha = 1 \end{cases} \tag{1}$$

Alpha fairness is achieved by maximizing $W_\alpha(\boldsymbol{u})$ subject to a limit on the number of individuals that can be selected.

We let binary variable $x_i = 1$ when individual i is selected. The expected utility gained by individual i is therefore $a_i x_i + b_i$. The social welfare resulting from a given vector $\boldsymbol{x} = (x_1, \ldots, x_n)$ of selection decisions, as measured by the alpha fairness SWF, is

$$W_\alpha(\boldsymbol{x}) = \begin{cases} \dfrac{1}{1-\alpha} \displaystyle\sum_{i=1}^n (a_i x_i + b_i)^{1-\alpha}, & \text{if } \alpha \geq 0 \text{ and } \alpha \neq 1 \\ \displaystyle\sum_{i=1}^n \log(a_i x + b_i), & \text{if } \alpha = 1 \end{cases} \tag{2}$$

If m $(< n)$ individuals are to be selected, one achieves alpha fairness for a given α by maximizing $W_\alpha(\boldsymbol{x})$ subject to $\sum_{i=1}^n x_i = m$. A maximizing vector $\boldsymbol{x}$ can be deduced using a simple greedy algorithm. We first consider the case $\alpha \neq 1$. The top expression in (2) can be written as

$$\frac{1}{1-\alpha} \sum_{i=1}^n b_i^{1-\alpha} + \frac{1}{1-\alpha} \sum_{i=1}^n \Big((a_i x_i + b_i)^{1-\alpha} - b_i^{1-\alpha} \Big) \tag{3}$$

Since the first term is a constant, we can maximize (3) by maximizing its second term, which can be written as

$$\frac{1}{1-\alpha} \sum_{i|x_i=1} \Big((a_i + b_i)^{1-\alpha} - b_i^{1-\alpha} \Big) = \sum_{i|x_i=1} \Delta_i(\alpha) \tag{4}$$

where we define

$$\Delta_i(\alpha) = \begin{cases} \frac{1}{1-\alpha} \big((a_i + b_i)^{1-\alpha} - b_i^{1-\alpha} \big), & \text{if } \alpha \geq 0,\ \alpha \neq 1 \\ \log(a_i + b_i) - \log(b_i), & \text{if } \alpha = 1 \end{cases}$$

The term $\Delta_i(\alpha)$ is the increase in welfare that results from selecting individual i (for a given $\alpha \neq 1$). We can maximize (4) subject to $\sum_{i=1}^n x_i = m$ by selecting the m individuals with the largest *welfare differential* $\Delta_i(\alpha)$. A similar argument applies for $\alpha = 1$. Thus we have

Theorem 1. *If $\Delta_{\pi_1}(\alpha) \geq \cdots \geq \Delta_{\pi_n}(\alpha)$, where $\pi_1, \ldots, \pi_n$ is a permutation of $1, \ldots, n$, then one can maximize $W_\alpha(\boldsymbol{x})$ subject to $\sum_{i=1}^n x_i = m$ by setting $x_i = 1$ for $i = \pi_1, \ldots, \pi_m$, and $x_i = 0$ for $i = \pi_{m+1}, \ldots, \pi_n$.*

At this point we can easily check whether achieving alpha fairness results in the various forms of group parity by observing whether their definitions are satisfied when individuals $\pi_1, \ldots, \pi_m$ are selected.

4 Modeling Protected and Nonprotected Groups

While Theorem 1 specifies an alpha fair selection policy for any given set of individual utility parameters $(a_1, b_1), \ldots (a_n, b_n)$, it yields limited insight into how the utility characteristics of protected and nonprotected groups affect alpha fair selections. In addition, the large number of parameters makes relationships difficult to analyze in a comprehensible fashion.

We address these issues by supposing that the expected utilities in the two groups occur on sliding scale. Specifically, we suppose that the selection benefits a_i in the nonprotected group are distributed uniformly on a scale from a maximum $A_{\max}$ down to a minimum $A_{\min}(< A_{\max})$, and selection benefits in the protected group vary uniformly from $a_{\max}$ down to $a_{\min}(< a_{\max})$. A nonuniform distribution is more realistic, but it requires a complicated analysis that is harder to interpret, while yielding basically the same qualitative results. To further simplify analysis, we suppose that the base utility has the same value B for all nonprotected individuals, and the same value b for all protected individuals. We assume that $B > b$ and, consistent with the previous section, that $A_{\min} + B > 0$ and $a_{\min} + b > 0$. Finally, we suppose that the protected group comprises a fraction β of the population, with $0 < \beta < 1$.

We further assume that individuals within a given group are selected in decreasing order of their selection benefit. Thus if a fraction S of nonprotected individuals are selected, the last individual selected in that group has the selection benefit $A(S) = (1 - S)A_{\max} + SA_{\min}$ and a social welfare differential of

$$\Delta_S(\alpha) = \begin{cases} \frac{1}{1-\alpha}\Big(\big(A(S) + B\big)^{1-\alpha} - B^{1-\alpha}\Big), & \text{if } \alpha \geq 0,\ \alpha \neq 1 \\ \log\big(A(S) + B\big) - \log(B), & \text{if } \alpha = 1 \end{cases}$$

Similarly, if a fraction s of individuals are selected in the protected group, the last individual selected has the selection benefit $a(s) = (1 - s)a_{\max} + sa_{\min}$ and the social welfare differential

$$\Delta'_s(\alpha) = \begin{cases} \frac{1}{1-\alpha}\Big(\big(a(s) + b\big)^{1-\alpha} - b^{1-\alpha}\Big), & \text{if } \alpha \geq 0,\ \alpha \neq 1 \\ \log\big(a(s) + b\big) - \log(b), & \text{if } \alpha = 1 \end{cases}$$

We will suppose that the population is large enough that S and s can be treated as continuous variables. This simplifies the analysis considerably without materially affecting the conclusions.

Since the social welfare differential is a monotone increasing function of the selection benefit, selecting individuals in order of decreasing welfare differential is, within each group, the same as selecting in order of decreasing selection benefit. By Theorem 1, selection in order of decreasing welfare differential maximizes the alpha fairness SWF subject to $\sum_i x_i = m$ if we select individuals until the desired fraction $\sigma = m/n$ of the population is selected. This occurs when

$$(1-\beta)S + \beta s = \sigma, \quad \text{or} \quad s = s(S) = \frac{\sigma - (1-\beta)S}{\beta} \tag{5}$$

We first take note of the ranges within which S and s can vary, subject to (5). Since we must have $0 \leq S \leq 1$ and $0 \leq s \leq 1$, S can vary in the range from $S_{\min}$ to $S_{\max}$, where

$$S_{\min} = \max\left\{0, \; \frac{\sigma - \beta}{1-\beta}\right\}, \quad S_{\max} = \min\left\{1, \; \frac{\sigma}{1-\beta}\right\}$$

and s can vary from $s(S_{\max})$ to $s(S_{\min})$. Now since $A_{\max} > A_{\min}$, $\Delta_S(\alpha)$ is monotone decreasing in S. Similarly, $\Delta'_s(\alpha)$ is monotone decreasing in s, so that $\Delta'_{s(S)}(\alpha)$ is monotone increasing in S. This means that we can consider three cases, illustrated by Fig. 1:

(a) $\Delta_{S_{\min}}(\alpha) > \Delta'_{s(S_{\min})}(\alpha)$ and $\Delta_{S_{\max}}(\alpha) \geq \Delta'_{s(S_{\max})}(\alpha)$.
(b) $\Delta_{S_{\min}}(\alpha) \leq \Delta'_{s(S_{\min})}(\alpha)$ and $\Delta_{S_{\max}}(\alpha) < \Delta'_{s(S_{\max})}(\alpha)$.
(c) $\Delta_{S_{\min}}(\alpha) > \Delta'_{s(S_{\min})}(\alpha)$ and $\Delta_{S_{\max}}(\alpha) < \Delta'_{s(S_{\max})}(\alpha)$

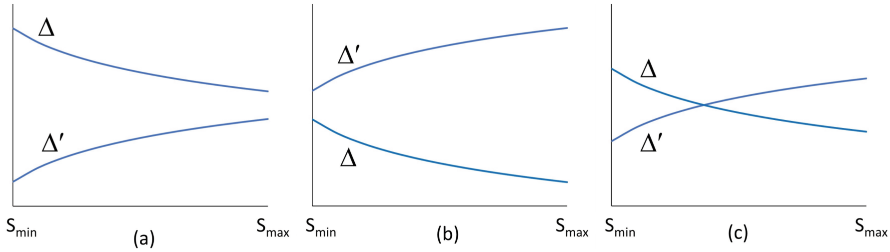

Fig. 1. Cases (a), (b), and (c) in the proof of Theorem 2

Theorem 2. *Suppose that individuals are selected in decreasing order of their selection benefit, and let S^* and $s^* = s(S^*)$, respectively, be the fraction of the nonprotected and protected groups selected at the end of the selection process.*

Then for a sufficiently large population, S^* and s^* achieve alpha fairness if and only if

$$\begin{cases} (S^*, s^*) = \left(\min\left\{1, \frac{\sigma}{1-\beta}\right\},\ \frac{\sigma}{\beta}\left[1 - \min\left\{1, \frac{1-\beta}{\sigma}\right\}\right]\right), & \textit{in case (a)} \\ (S^*, s^*) = \left(\frac{\sigma}{1-\beta}\left[1 - \min\left\{1, \frac{\beta}{\sigma}\right\}\right],\ \min\left\{1, \frac{\sigma}{\beta}\right\}\right), & \textit{in case (b)} \\ \Delta_{S^*}(\alpha) = \Delta'_{s(S^*)}(\alpha), & \textit{in case (c)} \end{cases} \tag{6}$$

Proof. Recall that by Theorem 1, alpha fairness is achieved by selecting individuals in decreasing order of their welfare differential until $S = s(S)$. We consider the three cases separately. (a) Because $\Delta_S(\alpha) \geq \Delta'_{s(S)}(\alpha)$ for all $S \in [S_{\min}, S_{\max}]$, we select entirely from the nonprotected group until it is exhausted, and then move to the protected group if necessary to select a fraction σ of the population. Thus we can set

$$S^* = \min\left\{S_{\max}, \frac{\sigma}{1-\beta}\right\} = \min\left\{1, \frac{\sigma}{1-\beta}\right\}$$

where the first equality is due to the fact that we must have $S^* = \sigma/(1-\beta)$ in order to select a fraction σ if the population if $\sigma \leq 1-\beta$, and the second equality is due to the definition of $S_{\max}$. The expression given in (6) for $s^* = s(S^*)$ follows directly from the definition of $s(S^*)$, and it is easily checked that $s_{\min} \leq s^* \leq s_{\max}$ using the definitions of $s_{\min}$ and $s_{\max}$. (b) The argument is very similar to that of the previous case. (c) In this case, some but not all individuals are selected in both groups. Let (S, s) be the fraction of the nonprotected and protected individuals selected at any given point in the selection process. We first show that $\Delta_S(\alpha) = \Delta'_s(\alpha)$ for a sufficiently large population. Let Δ_0 and Δ_1 be the welfare differentials of the last two nonprotected individuals selected, and Δ'_0 and Δ'_1 the differentials of the last two protected individuals selected. Their selection order is necessarily one of the following: $(\Delta_0, \Delta'_0, \Delta_1, \Delta'_1)$, $(\Delta_0, \Delta'_0, \Delta'_1, \Delta_1)$, $(\Delta'_0, \Delta_0, \Delta_1, \Delta'_1)$, $(\Delta'_0, \Delta_0, \Delta'_1, \Delta_1)$. In each case, $|\Delta_1 - \Delta'_1|$ is at most $\max\{\Delta_0 - \Delta_1, \Delta'_0 - \Delta'_1\}$. For a sufficiently large population, $\Delta_0 - \Delta_1$ and $\Delta'_0 - \Delta'_1$ are arbitrarily small, and so $|\Delta_1 - \Delta'_1|$ is arbitrarily small. Thus we have $\Delta_S(\alpha) = \Delta'_s(\alpha)$ throughout the selection process, and in particular at the end of the process, when $(S, s) = (S^*, s^*)$. The theorem follows. □

To explore how alpha fair selection policies depend on the utility characteristics of protected and nonprotected groups, we define three scenarios that represent qualitatively different practical situations.

Scenario 1. Protected individuals are somewhat less likely to benefit from being selected, as when those selected for job interviews are less likely to be hired due to less obvious qualifications. Here, $[A_{\min}, A_{\max}] = [0.5, 1.5]$ and $[a_{\min}, a_{\max}] = [0.2, 1]$.

Scenario 2. Some protected individuals can benefit more than anyone else from selection, as when talented but economically disadvantaged individuals are admitted to a university. Here, $[A_{\min}, A_{\max}] = [0.5, 0.8]$ and $[a_{\min}, a_{\max}] = [0.2, 1]$.

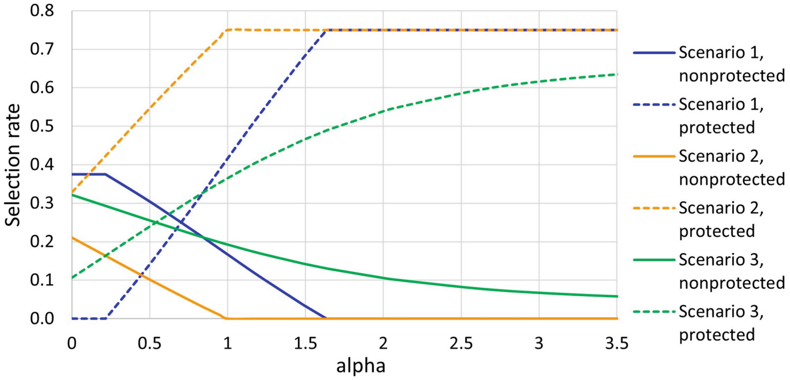

Fig. 2. Alpha fair selection rates, assuming overall selection rate of 0.25

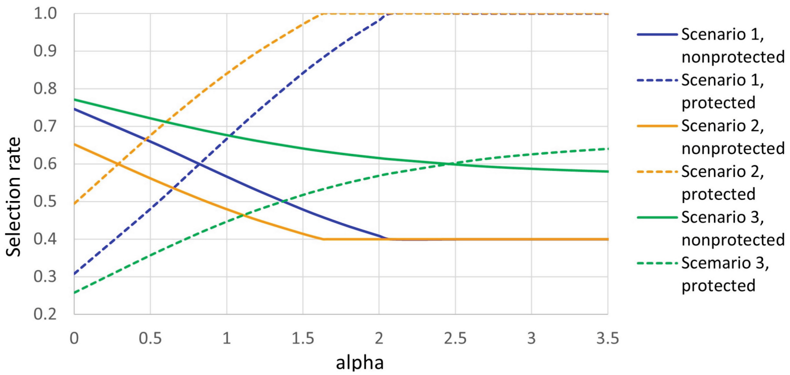

Fig. 3. Alpha fair selection rates, assuming overall selection rate of 0.6

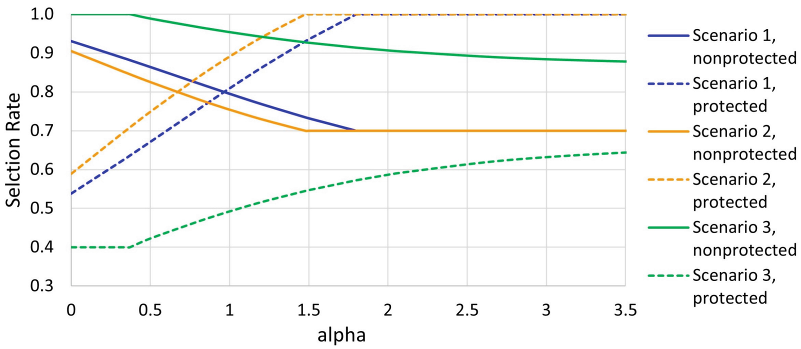

Fig. 4. Alpha fair selection rates, assuming overall selection rate of 0.8

Scenario 3. Significantly many protected individuals are likely to be harmed by selection, as when failure to repay a mortgage results in eviction. Here, $[A_{\min}, A_{\max}] = [0.5, 1]$ and $[a_{\min}, a_{\max}] = [-0.5, 1]$.

Plots of alpha fair selection policies in these scenarios appear in each of Figs. 2, 3 and 4. The three figures respectively assume overall selection rates of $\sigma = 0.25, 0.6, 0.8$. These selection rates are chosen to be less than, equal to, and greater than a qualification rate of 0.6, which will be assumed for subsequent plots of alpha fair odds ratios and predictive rates.

The plots show the relationship between alpha fair selection rates (S, s) and the chosen value of α. As expected, larger values of α (indicating a greater emphasis on fairness) result in higher selection rates in the protected group (dashed curves) and lower rates in the nonprotected group (solid curves). Scenario 1 calls for lower section rates in the protected group than Scenario 2 because of the greater utility cost of achieving fairness in Scenario 1; recall that alpha fairness consider total utility as well as Rawlsian fairness. Both scenarios require selecting the entire protected group for sufficiently large α, except when $\sigma = 0.25$, in which case the small number of selections does not exhaust the protected group. In Scenario 3, by contrast, the protected group's selection rate approaches 2/3 asymptotically, because only 2/3 of the group benefits from being selected in this scenario.

5 Demographic Parity

Demographic parity is achieved when $P(D|\neg Z) = P(D|Z)$. In the above model, this occurs when $s = S = \sigma$. As it turns out, cases (a) and (b) of Theorem 2 do not apply, and we can achieve demographic parity only by choosing a value of α (if one exists) dictated by case (c).

Theorem 3. *An alpha fair selection policy for a given α results in demographic parity if and only if there exists a selection rate S^* that satisfies the equation $\Delta_{S^*}(\alpha) = \Delta'_{S^*}(\alpha)$, in which case (S^*, S^*) is such a policy.*

Proof. We first note as follows that neither case (a) nor (b) in Theorem 2 applies. In case (a), demographic parity requires that

$$\min\left\{1, \frac{\sigma}{1-\beta}\right\} = \sigma, \quad \text{or} \quad \min\left\{\frac{1}{\sigma}, \frac{1}{1-\beta}\right\} = 1$$

This cannot hold, because $\beta > 0$ and $\sigma < 1$. Case (b) is similarly ruled out. We are therefore left with case (c), wherein Theorem 2 implies that $S^* = s(S^*)$ if and only if $\Delta_{S^*}(\alpha) = \Delta'_{S^*}(\alpha)$, as claimed. □

In Figs. 2, 3 and 4, demographic parity is achieved at the value of α where the rising and falling curves for a given scenario intersect. For example, if the overall section rate is $\sigma = 0.6$, parity is achieved in Scenario 1 when $\alpha = 0.833$ (Fig. 3). An important lesson in these plots is that a relatively small value of α frequently results in parity. That is, parity achieves a rather modest degree of fairness when utilities are taken into account. Indeed, proportional fairness ($\alpha = 1$), which is something of an industrial benchmark, typically calls for selecting a significantly greater fraction of the protected group than the nonprotected group. This is not the case in Scenario 3, however, where parity requires selecting protected individuals who receive minimal benefit and even harm from being selected. For example, no value of α corresponds to parity when $\sigma = 0.8$ (Fig. 4) because alpha fairness never endorses harmful choices.

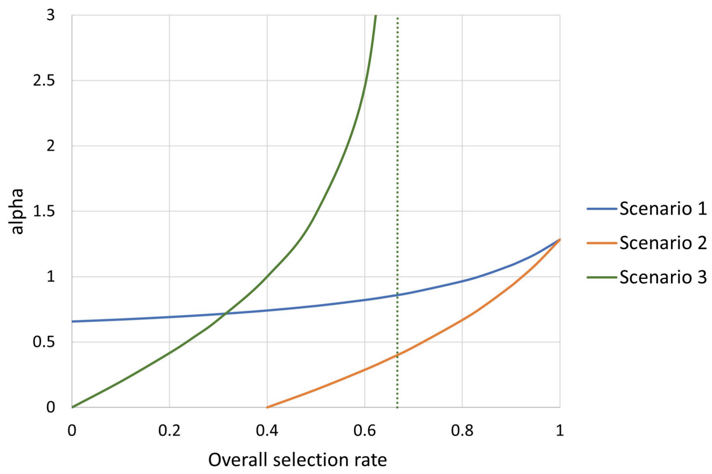

Fig. 5. Values of alpha that achieve demographic parity

Figure 5 provides a fuller picture of the relation between the selection rate σ and parity-achieving values of α. As σ increases, parity corresponds to larger values of α because it becomes necessary to select protected individuals who benefit little from selection. The curves for Scenarios 1 and 2 happen to meet at $\sigma = 1$ in this example because $A_{\min}$, $a_{\min}$, B, and b are the same in the two scenarios. We also note that $\alpha \to \infty$ as $\sigma \to 2/3$ in Scenario 3 because $\sigma > 2/3$ requires selecting individuals who are harmed by selection.

Interestingly, smaller values of α correspond to parity in Scenario 2 than in Scenario 1, despite the fact that rejection can be quite costly to some members of the protected group in Scenario 2 (due to their higher selection benefits). This occurs because a purely utilitarian assessment already takes this cost into account.

6 Equalized Odds

Equalized odds are achieved when $P(D|Y, Z) = P(D|Y, \neg Z)$. To define equalized odds in the above model, we suppose that a fraction Q of nonprotected individuals are qualified, and a fraction q of protected individuals are qualified. The a fraction $(1-\beta)Q+\beta q$ of the population is qualified. We also make the reasonable assumption that the selection benefit is greater for qualified individuals than unqualified individuals within a given group. Thus since $\Delta_S(\alpha)$ and $\Delta'_s(\alpha)$ are monotone decreasing as S and s increase, the qualified individuals in the nonprotected group consist of the fraction Q with the largest welfare differentials. The odds ratio for the nonprotected group is S/Q when $S \leq Q$ and 1 when $S > Q$, since in the latter case all the qualified individuals are selected. Thus the odds ratio is $\min\{1, S/Q\}$ for the nonprotected group, and similarly for the protected group. This means that we have equalized odds when

$$\min\left\{\frac{S}{Q}, 1\right\} = \min\left\{\frac{s}{q}, 1\right\}$$

This leads to the following theorem. It is convenient to define ρ to be the ratio of the fraction selected to the fraction of the population that is qualified, so that

$$\rho = \frac{\sigma}{(1-\beta)Q+\beta q}$$

Theorem 4. *An alpha fair selection policy $(S^*, s(S^*))$ for a given α results in equalized odds if and only if one of the following holds:*

$$S^* = Q\rho \leq Q \;\; \textit{and} \;\; s(S^*) = q\rho \leq q \tag{7}$$

$$S^* \geq Q \;\; \textit{and} \;\; s(S^*) \geq q \tag{8}$$

Proof. We consider four mutually exclusive and exhaustive cases:
(a) $S^* \leq Q$ and $s(S^*) \leq q$ (c) $S^* > Q$ and $s(S^*) \leq q$
(b) $S^* \leq Q$ and $s(S^*) > q$ (d) $S^* > Q$ and $s(S^*) > q$

In case (a), equalized odds is equivalent to $S^*(\alpha)/Q = s^*(\alpha)/q$, which implies $S^* = Q\rho$ and $s(S^*) = q\rho$ in (7) due to (5). Conversely, we can see as follows that either of the conditions (7) and (8) implies equalized odds. Under condition (7), the values for S^* and $s(S^*)$ in (7) imply $S^*/Q = s(S^*)/q$, and we have equalized odds. Under condition (8), both odds ratios are 1, and we again have equalized odds. In case (b), equalized odds implies $S^* = Q$, in which case condition (8) is satisfied. Conversely, the case hypothesis is consistent with only condition (8), in which case both odds ratios are 1 and we have equalized odds. Case (c) is similar. In case (d), one of the conditions (7)–(8) is necessarily satisfied (because the latter is satisfied), and we necessarily have equalized odds, because both odds ratios are 1. □

To continue the example of the previous section, we suppose that the qualification rates are $(Q, q) = (0.65, 0.5)$, so that a fraction 0.6 of the population is qualified. Figures 6 and 7, corresponding to $\sigma = 0.25$ and $\sigma = 0.6$, show alpha fair odds ratios for various α. No plot is given for $\sigma = 0.8$ because nearly all of the odds ratios are 1 due to the fact that considerably more individuals are selected than are qualified. In the important special case where the number selected is equal to the number qualified (Fig. 7), equalized odds is achieved only by an

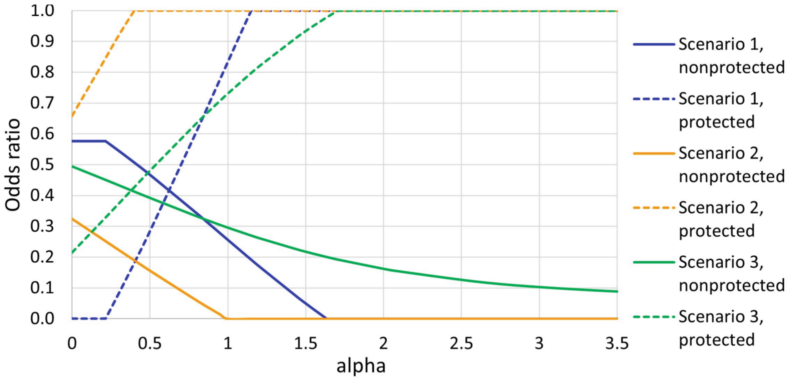

Fig. 6. Alpha fair odds ratios, assuming overall selection rate of 0.25.

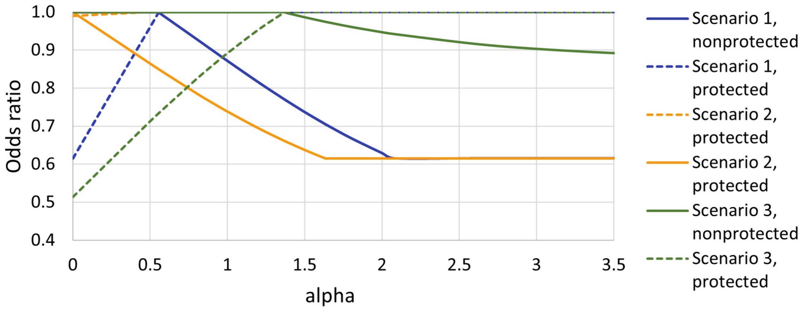

Fig. 7. Alpha fair odds ratios, assuming overall selection rate of 0.6.

accuracy-maximizing solution: precisely the qualified individuals are selected in both groups. This rules out any adjustment for fairness. The odds ratio is perhaps more useful when limited resources compel one to reject significantly many qualified individuals. In this event, somewhat smaller values of α are typically necessary to achieve equalized odds than demographic parity (Fig. 6). In Scenario 2, a purely utilitarian solution already achieves a higher odds ratio for the protected group, since some of its qualified members derive more utility from selection than anyone in the nonprotected group.

7 Predictive Rate Parity

Predictive rate parity is achieved when $P(Y|D,Z) = P(Y|D,\neg Z)$. The predictive rate for the nonprotected group is Q/S when $S \geq Q$ and 1 when $S < Q$, since in the latter case all the selected individuals are qualified. Thus the predictive rate is $\min\{Q/S, 1\}$ for the nonprotected group, and similarly for the protected group. This means that we have predictive rate parity when

$$\min\left\{\frac{Q}{S},1\right\} = \min\left\{\frac{q}{s},1\right\}$$

This leads to the following theorem, whose proof is very similar to the proof of Theorem 4.

Theorem 5. *An alpha fair selection policy* $(S^*, s(S^*))$ *for a given* α *results in predictive rate parity if and only if one of the following holds:*

$$S^* = Q\rho \geq Q \;\; and \;\; s(S^*) = q\rho \geq q \tag{9}$$

$$S^* \leq Q \;\; and \;\; s(S^*) \leq q \tag{10}$$

Note that the expressions for S^* *and* $s(S^*)$ *in* (9) *are the same as in* (7).

Figures 8 and 9, corresponding to $\sigma = 0.6$ and $\sigma = 0.8$, show alpha fair predictive rates for various α. There is no plot for $\sigma = 0.25$, because nearly all of the predictive rates are 1. We also note that larger predictive rates correspond to *smaller* values of α.

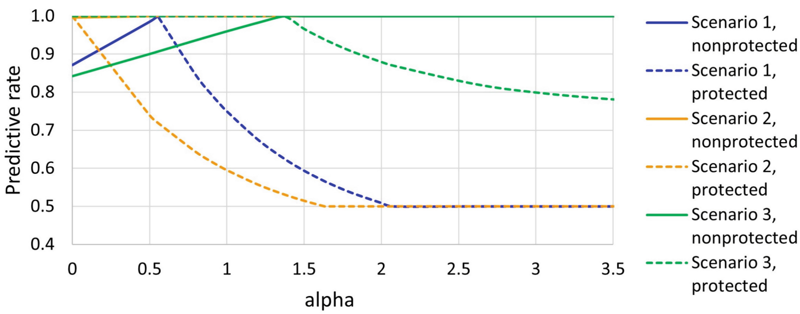

Fig. 8. Alpha fair predictive rates, assuming overall selection rate of 0.6.

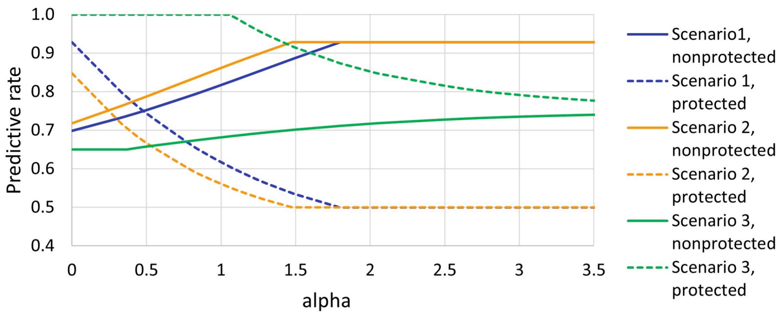

Fig. 9. Alpha fair predictive rates, assuming overall selection rate of 0.8.

8 Conclusion

Our aim in this paper has been to explore the extent to which social welfare optimization can assess well-known statistical parity metrics as criteria for group fairness in AI. Our focus on alpha fairness allows us to address parity questions by appealing to a well-studied concept of just distribution with theoretical underpinning. We conclude in this section by recalling the two problems associated with parity metrics and summarizing how they might be addressed from an optimization perspective.

1. *Accounting for welfare.* The alpha fairness criterion allows us to take explicit account of welfare implications, for various levels of fairness as indicated by the α parameter. We find that for certain values of α and certain group characteristics, an alpha fair selection policy can result in group parity of any of the three types. Yet it can also call for significant statistical *disparity* in order to achieve an acceptable distribution of utilities.

In particular, the alpha values that result in parity typically lie significantly below that corresponding to proportional fairness ($\alpha = 1$)—except when some individuals in the protected group are actually harmed by being selected, in which case larger values of α correspond to parity. Since proportional fairness is the most widely defended and applied variety of alpha fairness, it is noteworthy that it often requires, not parity, but *higher* selection rates for the protected group than for the rest of the population. In addition, a lower level of fairness (i.e., a smaller α) is necessary to achieve parity when rejection is more costly to members of the protected group than the rest of the population, other things being equal. This is because even a purely utilitarian accounting already takes this cost into account.

2. *Selecting and justifying parity metrics.* We derive a number of conclusions regarding the choice of parity metric. In general, we find that the implications of alpha fairness depend heavily on how many individuals are selected relative to the total number qualified, at least where equalized odds and predictive rate parity are concerned.

To elaborate on this, we first suppose that the total number selected is the same (or approximately the same) as the total number who are qualified in the population as a whole. In this case, demographic parity follows the pattern described above, in which relatively small values of α result in parity, except when some protected individuals are harmed by selection. Yet equalized odds, as well as predictive rate parity, are achieved if and only if the odds ratios and the predictive rates are 1 in both groups. This corresponds to an accuracy maximizing policy of selecting all and only qualified individuals. As a result, neither equalized odds nor predictive rate parity reflects any consideration of fairness beyond mere accuracy, and consequently neither is suitable as a fairness criterion in this context.

We next suppose that the total number selected is significantly less than the number qualified, presumably a common situation due to limited resources. In this case, equalized odds is generally achieved for smaller values α than are required for demographic parity, considerably smaller when some protected individuals are harmed by selection. This indicates that demographic parity demands a greater emphasis on fairness than equalized odds. This is consonant with the fact that equalized odds is sometimes seen as more easily defended, perhaps on grounds of equality of opportunity, than is demographic parity, which may reflect a desire to compensate for historically unjust discrimination. As for the predictive rate, it is almost always 1 when a significant number of qualified individuals are rejected, since those who make it through the sieve are almost always qualified. This means that predictive rate parity is likely to be achieved simply due to the high rejection rate and is therefore of little value as a fairness criterion.

Finally, we suppose that the number selected is significantly greater than the number qualified. Here, the odds ratio loses interest because it is almost always 1. While predictive rate parity becomes meaningful in this case, decision makers may be reluctant to select more individuals than are qualified. To the extent this is true, predictive rate parity has limited usefulness. A possible exception arises in the controversy over parole mentioned earlier. Predictive rate parity might be defended on the ground that a lower recidivism rate in the protected group (and therefore a higher predictive rate) may reflect stricter parole criteria than for other inmates [2]. Greater fairness may therefore require a *reduction* in the predictive rate of the protected group, which we have seen can be achieved by choosing a larger value of α. If this is taken as justifying a practice of paroling more individuals than are qualified (perhaps in order to reduce the predictive rate of protected individuals without tightening the criteria for others), then predictive rate parity could be a suitable criterion.

In summary, demographic parity can under certain conditions correspond to an alpha fair policy, but it may result in less fairness than desired for the protected group. Equalized odds can be a useful criterion when fewer individuals are selected than are qualified to be selected, but it corresponds to an even lesser degree of fairness. Predictive parity is a meaningful fairness measure only in the perhaps rather uncommon situation when decision makers select significantly more individuals than are qualified.

The foregoing conclusions regarding equalized odds and predictive rate parity rest on the assumption that, *within a given group*, qualified individuals are selected before unqualified individuals. This assumption might be defended on the ground that (a) qualified individuals are likely to benefit more from being selected, and (b) individuals who benefit more from being selected are selected first in the group. Assumption (a) might be based on observations that less qualified individuals pose a greater risk of defaulting on a mortgage, failing to secure a job, committing a crime while on parole, and so forth, and therefore have less expected benefit. As for (b), there is no apparent rationale, based on either expected utility or fairness, for selecting individuals within a group in any other order. It therefore seems reasonable to suppose (b) is true before assessing fairness.

We believe these results suggest that there is potential in an optimization perspective to inform fairness debates in AI. Further research could explore the parity implications of alternative social welfare functions, such as the Kalai-Smorodinsky and threshold criteria cited earlier. A particularly interesting research issue is the extent to which achieving fairness in the population as a whole can result in a reasonable degree of parity across all groups. This would obviate the necessity of selecting which groups to regard as protected, and how to balance their interests.

References

1. Angwin, J., Larson, J., Mattu, S., Kirchner, L.: Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica (2016). Accessed 23 May 2016
2. Anwar, S., Fang, H.: Testing for racial prejudice in the parole board release process: theory and evidence. J. Legal Stud. **44**, 1–37 (2015)
3. Barocas, S., Hardt, M., Narayanan, A.: Fairness and Machine Learning: Limitations and Opportunities. MIT Press, Cambridge (2023)
4. Baumann, J., Hannó, A., Heitz, C.: Enforcing group fairness in algorithmic decision making: utility maximization under sufficiency. In: Proceedings of FAccT 2022 (2022)
5. Bertsimas, D., Farias, V., Trichakis, N.: On the fairness-efficiency trade-off. Manag. Sci. **58**, 2234–2250 (2012)
6. Binmore, K., Rubinstein, A., Wolinsky, A.: The Nash bargaining solution in economic modelling. RAND J. Econ. 176–188 (1986)
7. Binns, R.: Fairness in machine learning: lessons from political philosophy. Proc. Mach. Learn. Res. **8**, 1–11 (2018)
8. Card, D., Smith, N.: On consequentialism and fairness. Front. Artif. Intell. **3**, 34 (2020)
9. Carter, I., Page, O.: When is equality basic. Aust. J. Philos. **101**, 983–997 (2022)
10. Castelnovo, A., Crupi, R., Greco, G., Regoli, D., Penco, I.G., Cosentini, A.C.: A clarification of the nuances in the fairness metrics landscape. Sci. Rep. **12**, 4209 (2022)
11. Chen, V., Hooker, J.N.: A just approach balancing Rawlsian leximax fairness and utilitarianism. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pp. 221–227 (2020)
12. Chen, V., Hooker, J.N.: Combining leximax fairness and efficiency in an optimization model. Eur. J. Oper. Res. **299**, 235–248 (2022)
13. Chen, V., Hooker, J.N.: A guide to formulating fairness in an optimization model. Ann. Oper. Res. **326**, 581–619 (2023)
14. Chouldechova, A.: Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. Big Data **5**(2), 153–163 (2017)
15. Corbett-Davies, S., Gaebler, J.D., Nilforoshan, H., Shroff, R., Goel, S.: The measure and mismeasure of fairness: a critical review of fair machine learning. J. Mach. Learn. Res. **24**, 1–117 (2023)
16. Dieterich, W., Mendoza, C., Brennan, T.: COMPAS risk scales: Demonstrating accuracy equity and predictive parity. Report , Northpointe Inc., Research Department (2016)
17. Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., Venkatasubramanian, S.: Certifying and removing disparate impact. In: Proceedings of 21st SIGKDD. ACM (2017)

18. Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., Srikumar, M.: Principled artificial intelligence: mapping consensus in ethical and rights-based approaches to principles for AI. Berkman Klein Center Research Publication No. 2020-1 (2020)
19. Friedler, S.A., Scheidegger, C., Venkatasubramanian, S.: On the (im)possibility of fairness. Commun. ACM **64**, 136–143 (2021)
20. Greene, J.: Moral Tribes: Emotion, Reason, and the Gap between Us and Them. Penguin Press, London (2013)
21. Harsanyi, J.C.: Rational Behaviour and Bargaining Equilibrium in Games and Social Situations. Cambridge University Press, Cambridge (1977)
22. Hooker, J.N., Williams, H.P.: Combining equity and utilitarianism in a mathematical programming model. Manag. Sci. **58**, 1682–1693 (2012)
23. Hu, L., Chen, Y.: Welfare and distributional impacts of fair classification (2018). arXiv preprint arXiv:1807.01134
24. Hu, L., Chen, Y.: Fair classification and social welfare. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 535–545 (2020)
25. Jobin, A., Ienca, M., Vayena, E.: The global landscape of AI ethics guidelines. Nat. Mach. Intell. **1**, 389–399 (2019)
26. Kalai, E., Smorodinsky, M.: Other solutions to Nash's bargaining problem. Econometrica **43**, 513–518 (1975)
27. Karsu, O., Morton, A.: Inequality averse optimization in operational research. Eur. J. Oper. Res. **245**, 343–359 (2015)
28. Kelly, F.P., Maulloo, A.K., Tan, D.K.H.: Rate control for communication networks: shadow prices, proportional fairness and stability. J. Oper. Res. Soc. **49**(3), 237–252 (1998)
29. Kleinberg, J., Mullainathan, S., Raghavan, M.: Inherent trade-offs in the fair determination of risk scores. In: Proceedings, Innovations in Theoretical Computer Science (ITCS). Dagstuhl Publishing, Germany (2017)
30. Lan, T., Chiang, M.: An axiomatic theory of fairness in resource allocation. Technical report. Princeton University (2011)
31. Lan, T., Kao, D., Chiang, M., Sabharwal, A.: An axiomatic theory of fairness in network resource allocation. In: Proceedings of the 29th Conference on Information communications (INFOCOM), pp. 1343–1351 (2010)
32. Leben, D.: Normative principles for evaluating fairness in machine learning. In: Proceedings, AAAI/ACM Conference on AI, Ethics, and Society, pp. 86–92 (2020)
33. Loi, M., Herlitz, A., Heidari, H.: A philosophical theory of fairness for prediction-based decisions. SSRN Electron. J. (2019)
34. Mazumdar, R., Mason, L., Douligeris, C.: Fairness in network optimal flow control: optimality of product forms. IEEE Trans. Commun. **39**(5), 775–782 (1991)
35. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Comput. Surv. (CSUR) **54**(6), 1–35 (2021)
36. Mo, J., Walrand, J.: Fair end-to-end window-based congestion control. IEEE/ACM Trans. Network. **8**, 556–567 (2000)
37. Moss, J.: How to value equality. Philos. Compass **10**, 187–196 (2015)
38. Nash, J.F.: The bargaining problem. Econometrica **18**, 155–162 (1950)
39. Ogryczak, W., Luss, H., Pióro, M., Nace, D., Tomaszewski, A.: Fair optimization and networks: a survey. J. Appl. Math. **2014**, 1–25 (2014)
40. Ogryczak, W., Wierzbicki, A., Milewski, M.: A multi-criteria approach to fair and efficient bandwidth allocation. Omega **36**(3), 451–463 (2008)
41. Rubinstein, A.: Perfect equilibrium in a bargaining model. In: Econometrica, pp. 97–109 (1982)

42. Selbst, A., Barocas, S.: Big data's disparate impact. Calif. Law Rev. **671**, 671–732 (2016)
43. Verloop, I.M., Ayesta, U., Borst, S.: Monotonicity properties for multi-class queueing systems. Disc. Event Dyn. Syst. **20**, 473–509 (2010)
44. Williams, A., Cookson, R.: Equity in Health. In: Culyer, A.J., Newhouse, J.P. (eds.) Handbook of Health Economics (2000)